Studies in Brain and Mind 18

Robert W. Clowes
Klaus Gärtner
Inês Hipólito *Editors*

The Mind-Technology Problem

Investigating Minds, Selves and 21st Century Artefacts

Springer

Studies in Brain and Mind

Volume 18

The series Studies in Brain and Mind provides a forum for philosophers and neuroscientists to discuss theoretical, foundational, methodological, and ethical aspects of neuroscience. It covers the following areas:

- Philosophy of Mind
- Philosophy of Neuroscience
- Philosophy of Psychology
- Philosophy of Psychiatry and Psychopathology
- Neurophilosophy
- Neuroethics

The series aims for a high level of clarity, rigor, novelty, and scientific competence. Book proposals and complete manuscripts of 200 or more pages are welcome. Original monographs will be peer reviewed. Edited volumes and conference proceedings will be considered provided that the chapters are individually refereed.

This book series is indexed in SCOPUS.

Initial proposals can be sent to the Editor-in-Chief, prof. Gualtiero Piccinini, at piccininig@umsl.edu. Proposals should include:

- A short synopsis of the work or the introduction chapter
- The proposed Table of Contents
- The CV of the lead author(s)
- If available: one sample chapter

We aim to make a first decision within 1 month of submission. In case of a positive first decision the work will be provisionally contracted: the final decision about publication will depend upon the result of the anonymous peer review of the complete manuscript. We aim to have the complete work peer-reviewed within 3 months of submission.

For more information, please contact the Series Editor at piccininig@umsl.edu.

More information about this series at http://www.springer.com/series/6540

Robert W. Clowes • Klaus Gärtner
Inês Hipólito
Editors

The Mind-Technology Problem

Investigating Minds, Selves and 21st Century Artefacts

Springer

Editors
Robert W. Clowes
Instituto de Filosofia da Nova (IFILNOVA)
Faculdade de Ciências Sociais e Humanas
Universidade Nova de Lisboa
Lisboa, Portugal

Inês Hipólito
Department of Philosophy & Berlin School of Mind and Brain
Humboldt-Universität zu Berlin
Berlin, Germany

Klaus Gärtner
Departamento de História e Filosofia das Ciências & Centro de Filosofia das Ciências da Universidade de Lisboa (CFCUL)
Faculdade de Ciências
Universidade de Lisboa
Lisboa, Portugal

ISSN 1573-4536 ISSN 2468-399X (electronic)
Studies in Brain and Mind
ISBN 978-3-030-72643-0 ISBN 978-3-030-72644-7 (eBook)
https://doi.org/10.1007/978-3-030-72644-7

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Acknowledgments

We would like to thank Joe Corabi, Richard Heersmink, Paul Smart, Dina Mendonça, Steve Fuller, Corey Maley, Gloria Andrada, João L. Cordovil, Jenny Imhoff, Gil Santos, Ian Robertson, Giovanna Colombetti, Wim De Neys, Henry Shevlin, Caio Novaes, the Lisbon Mind and Reasoning Group, the ArGLAB, the Instituto de Filosofia da NOVA, the Centro de Filosofia das Ciências da Universidade de Lisboa, the Universidade Nova de Lisboa, the Faculdade de Ciências Sociais e Humanas, the Universidade de Lisboa, the Faculdade de Ciências da Universidade de Lisboa, the Departamento de História e Filosofia das Ciências da Faculdade de Ciências da Universidade de Lisboa, the Fundação para a Ciência e a Tecnologia, the Fundação Luso-Americana para o Desenvolvimento Fciências.ID, and the participants and audience of the "Minds, Selves and 21st Century Technology" conference in Lisbon, Portugal, June 2016.

Special thanks to Gualtiero Piccinini.

We would like to extend our deepest gratitude to Susan Schneider. Without her effort towards the "Minds, Selves and 21st Century Technology" conference in Lisbon, Portugal, June 2016, and her ideas and contributions, this volume would not have been possible.

Robert W. Clowes's work is supported by FCT, "Fundação para a Ciência e a Tecnologia, I.P.'" by the Stimulus of Scientific Employment grant (DL 57/2016/CP1453/CT0021) and personal grant (SFRH/BPD/70440/2010). It has also been heavily supported by his family and Robert would like to express his gratitude to Susana and Rafael.

Klaus Gärtner's work is endorsed by the financial support of FCT, "Fundação para a Ciência e a Tecnologia, I.P." under the Stimulus of Scientific Employment (DL 57/2016/CP1479/CT0081) and by the Centro de Filosofia das Ciências da Universidade de Lisboa (UIDB/00678/2020). Further, Klaus would like thank his family, Ana and Elias, for their lasting support.

This work is endorsed by the FCT project "Emergence in the Natural Sciences: Towards a New Paradigm" (PTDC/FER-HFC/30665/2017).

Contents

About the Editors

Robert W. Clowes is senior researcher and coordinator of the Lisbon Mind and Reasoning Group at the Universidade Nova de Lisboa, Portugal. His research interests span a range of topics in philosophy and cognitive science, including the philosophy of technology, memory, agency, skills, and the implications of embodiment and cognitive extension for our understanding of the mind and conscious experience. He is particularly interested in the philosophical and cognitive scientific significance of new technologies, especially those involving the Internet and artificial intelligence and how these interact with human agency. His work has appeared in a variety of journals, including *TOPOI*, *Review of Philosophy and Psychology*, *AI & Society*, *Phenomenology and the Cognitive Sciences*, *Philosophy and Technology*, and the *Journal of Consciousness Studies*. He received his PhD from the University of Sussex.

Klaus Gärtner studied philosophy at the University of Regensburg. He obtained his PhD at the Instituto da Filosofia da NOVA (Universidade Nova de Lisboa). Currently, he is a researcher at the Departamento de História e Filosofia das Ciências and member of the Centro de Filosofia das Ciências da Universidade de Lisboa in the Faculdade de Ciências da Universidade de Lisboa. He is also a founding member of the Lisbon Mind and Reasoning Group. His research interests include philosophy of mind and cognitive science, philosophy of science, epistemology, and metaphysics.

Inês Hipólito is a postdoctoral fellow and a lecturer at the Berlin School of Mind and Brain (Humboldt Universität zu Berlin), and an affiliated member to the Neurobiology group at the Wellcome Centre for Human Neuroimaging (University College London). She works on the intersection between philosophy of cognition and computational neuroscience. More precisely, Hipólito applies tools of conceptual modelling to answer philosophical questions of cognition that are compatible with the formalisms of dynamical systems theory. Hipólito has co-edited special issues for *Philosophical Transactions*, *Consciousness and Cognition*, and the Mind and Brain Studies (Springer). She has published work in edited books (Routledge,

CUP) and journals (*Australasian Philosophical Review*, *Physics of Life Reviews*, *Progress in Biophysics and Molecular Biology*, *Synthese*, *Network Neuroscience*). Dr. Hipólito's work has been honored with international prizes and awards, including the Portuguese Ministry for Science and Higher Education; the University of Oxford; the Federation of European Neuroscience Societies; and an award by the British Association for Cognitive Neuroscience.

Chapter 1
The Mind Technology Problem and the Deep History of Mind Design

Robert W. Clowes, Klaus Gärtner, and Inês Hipólito

1.1 What Is the Mind Technology Problem?

We are living through a new phase in human development where much of everyday life – at least in the most technologically developed parts of the world – has come to depend upon our interaction with "smart" artefacts. Alongside this increasing adoption and ever-deepening reliance on intelligent machines, important changes have taken place, often in the background, as to how we think of ourselves and how we conceptualize our relationship with technology. As we design, create, and learn to live with a new order of artefacts which exhibit behavior that, were it to be carried out by human beings would be seen as intelligent, the ways in which we conceptualize intelligence, minds, reasoning and related notions such as self and agency are undergoing profound shifts. Indeed, it is possible to argue that the basic background assumptions informing, and the underlying conceptual scheme structuring our reasoning about minds has recently been transformed. This shift has changed the nature and quality of both our folk understanding of mind, our scientific psychology, and the philosophical problems that the interaction of these realms produce. Many of the traditional problems in the philosophy of mind have become reconfigured in the

R. W. Clowes (✉)
Instituto de Filosofia da Nova (IFILNOVA), Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa, Lisbon, Portugal
e-mail: robertclowes@fcsh.unl.pt

K. Gärtner
Departamento de História e Filosofia das Ciências & Centro de Filosofia das Ciências da Universidade de Lisboa (CFCUL), Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal
e-mail: kgartner@fc.ul.pt

I. Hipólito
Department of Philosophy & Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Berlin, Germany
e-mail: ines.hipolito@hu-berlin.de

R. W. Clowes et al. (eds.), *The Mind-Technology Problem*, Studies in Brain and Mind 18, https://doi.org/10.1007/978-3-030-72644-7_1

process. This book treats this reconfiguration of our concepts of mind and of technology, and the philosophical problems this reconfiguration engenders.

These new conceptualizations – sometimes implicit, sometimes explicit – about the nature of mind and its relationships to the artefacts we build has given rise to a new constellation of basic philosophical problems about the very nature of mind. This constellation we call, *The Mind-Technology Problem*. The mind-technology problem should be understood as the successor to the mind-body problem. The mind-body problem as we know it today has been developed, or perhaps clearly noticed and articulated, in response to the set of ideas formulated by Descartes in the seventeenth century. The problem – really it is better understood as a constellation of problems – seems to have emerged from conceptual incongruities generated by a change in the background or implicit metaphysics of the age, especially moves toward the new mechanistic philosophy (Wootton 2015).

Descartes's ideas arose in the first era of the mechanistic revolution in the seventeenth century, and the "mechanistic philosophy" that Descartes then championed. The idea that minds can be understood as mechanisms or can be explained by mechanistic processes is very recent (Boden 2006), and at least for the contemporaries of Descartes, was highly counter-intuitive if not quite impossible to conceive.[1] However, we have to carefully distinguish the notion that human beings are can be understood as machines from the idea that minds can be understood mechanistically. Boden credits Descartes with the emergence of the idea of 'man as machine´ as a major theme in experimental science. She notes however that Descartes's general scientific approach was mechanist in two different ways. "On the one hand, he believed that the principles of physics can explain all the properties of material things. On the other hand, he often drew explicit analogies between living creatures and man-made machines, seeing these as different in their complexity rather than their fundamental nature." (Boden 2006, p. 58).

Famously of course, Descartes did not extend this mechanist account to the explanation of the human mind. The mind-body problem as we know it today proceeds from the dualist assumption that mind is something essentially different from material stuff. According to this view, the mind cannot be embodied in, or realized by, supervene upon, or otherwise be imminent in the causal properties of matter. This is because mind itself is conceived of as a separate substance with its own

[1] In May 1643 Princess Elisabeth of Bohemia wrote to Descartes whose work she had been following closely and posed the problem of interaction in an especially pointed way. She asked, "how the mind of human being, being only a thinking substance, can determine the bodily spirits in producing bodily actions." Princess Elisabeth, pushing upon the central problem arising from substance dualism, found herself unsatisfied with Descartes's attempts to resolves the question to her satisfaction. In exasperation she finally writes "it would be easier for me to concede matter and extension to the mind than it would be for me to concede the capacity to move a body and be moved by one to an immaterial thing." Cited in Jaegwon Kim's Philosophy of Mind (Kim 2006, pp. 41–42). Kim notes that this is a (very) early example of the causal argument from materialism, that holds that mental causation implies materialism, for, it is hard to see how any putative immaterial substance might interact with the rest of the causal order. This is also a compelling incidence of how it is sometimes possible to think against the grain of even a highly dominant conceptual scheme.

special properties. This conceptual scheme creates the central problem for substance dualism, namely, the problem of interaction, i.e., how it is that being essentially different substances minds and matter can interact with each other at all. Ancillary problems such as the problem of other minds, mental causation or free will are related but configured around this central problem. This is what is meant by calling the mind/body problem a constellation problem. The constellation arranges philosophical problems from a particular vantage point that appears fixed.

It is true, there have been deep controversies around whether the mind-body problem is really one problem at all, or rather a series of problems. This is a highly contested matter, but it is interesting to note that in the general introduction to a recent six volume *History of the Philosophy of Mind* (Copenhaver and Shields 2019a, b) the series authors observe how it is neither clear that the mind-body problem was clearly formulated by the ancients, nor by others before Descartes, nor that the mind-body problem has been construed in a consistent way since.[2] If the mind-body problem is *really* a constellation problem, then it appears that the constellation's configuration has undergone changes over the years with new problems being added and some falling out.[3] Yet, despite these uncertainties over the construal (or creation) of the mind-body problem as we know it today, it is generally agreed to fall out of a particular epoch of thought in the seventeenth century, and its reach and influence over how we continue to think of the mind, even today, are seldom disputed.

With a few notable exceptions, by far the majority view of the seventeenth century was that the mind and body were irreconcilably different substances.[4] However, substance Dualists tend to be thin on the ground these days, at least in philosophical and scientific circles.[5] If we are to look for the major reason for this change, it is not in the development and pursuance of philosophical arguments, but through developments in science – and of special interest here – technology. These developments have progressively made the dualist conceptual scheme more difficult to maintain. Substance dualism has become undoubtedly less conceptually lucid against the background of the information revolution and the computational metaphor for the mind. This is not to say that all problems in the philosophy of mind or even the mind-body problem have been resolved. Far from it. But at least for those working

[2] Indeed, it is only since the 1960s that there has been – at least in the Anglo-Saxon world – university courses which are explicitly targeted at philosophy of mind. Many of such courses are organized around the Mind Body problem.

[3] The sense that the mind and body are distinct has arguably been part of folk-psychology and religious views of the world for centuries, as well as metaphysical views from Plato to Descartes. That the mind, or the soul, is separate from matter was something that seems to be introduced only at a time when mechanist views of the rest of nature are being clearly articulated for the first time.

[4] Amy Kind for example argues that dualism was much the preferred view of the early modern period and materialist and what we would now call physicalist positions were much out of favour (Kind 2018). La Mettrie's ([1747] *Man a Machine*, was very much against the tide of ideas of the time although it anticipated major theme of twentieth century philosophy.

[5] Chalmers informational dualism is a notable exception here (Chalmers 2002). Arguably, in popular culture and in folk psychology a form of dualism is widespread.

in the contemporary philosophy and cognitive sciences, our understanding of the minds is generally understood against a far different conceptional background to that trailblazed by Descartes. This background is the computational or informational conception of mind and mental processes.

An argument could be made that there is no singular conceptual scheme for mind anymore but rather a series of overlapping and often rather contradictory frameworks that the folk use to conceptualize their minds and cognitive processes. At the same time, the constellation of philosophical problems we face when accounting for minds seems to have undergone a profound shift as new computationally inflected conceptual models have arisen. It is true, as Daniel Dennett (1991) has famously argued, that a deep Cartesian influence remains in the conceptual backdrop of many otherwise materialist theories of mind in the form of what he calls "The Cartesian Theatre". For Dennett, any view that holds that there is some place in the brain or consciousness where it "all comes together" is implicitly Cartesian even if the proponents of such a view hold themselves to be thoroughgoing materialists. There also seems to be a minor industry in philosophy, at least from Ryle (1949) onwards, pointing out various implicit dualisms and how they continue to contaminate the contemporary sciences of mind – and the works of other philosophers. Yet, widely held conceptual schemes such as those that underlie folk psychology may be highly internally heterogeneous and relatively immune from problems of contradiction, at least in the short term. Elements of substance dualism, cognitive psychology, Freudian psychoanalysis alongside the computational model of mind seem to enjoy an uneasy co-existence in the contemporary folk understanding of the mental. Nevertheless, over the last 70 years, computationalism seems to have fundamentally reshaped many of our concepts and categories for thinking about minds.

The idea that a background – and technologically influenced – conceptual scheme shapes our arguments and abilities to form inferences is perhaps not given enough consideration in analytic philosophy or psychology.[6] However, there are some notable accounts which take these constraints much more seriously. A central reference point here is MIT history professor Bruce Mazlish (1993) book *The Fourth Discontinuity: The Co-Evolution of Humans and Machines*. Mazlish's book builds upon an idea, originally suggested by Sigmund Freud, that the history of human self-conceptualizations in the Western Tradition have developed through a series of discontinuities or shocks to our sense of ourselves and place in the universe. Against the background of the Judeo-Cristian idea that Man – today we would say human beings – is central to Creation and the universe with him at the

[6] An important exception to this generalization is Richard Gregory's monumental (1981) *Mind in Science: A History of Explanations in Psychology.* Important work on how our conceptual schemes are more generally constrained by technology and the history of invention can be found in Postman (1993). One field where the background metaphors for mind are considered is cognitive linguistics (Fauconnier and Turner 2002; Lakoff and Johnson 2003 [1980]), and perhaps especially in Lakoff and Johnson (1999). However even cognitive linguists tend to chiefly pay attention to the way that concepts are shaped by the nature of human embodiment. The idea that our use of technology may similarly shape our abstract reasoning about the nature of mind is less explored.

center of it (See Theiner, this volume), the role accorded to human beings, Freud claimed, has had to undergo a series of intellectual shocks. These shocks have both decentered us – literally, the human race appearing as ever less central to the universe – and at the same time forced us to rethink what, if anything, is so special about being human. The first conceptual shock was the proposal by Nicolaus Copernicus (1473–1543) of the heliocentric cosmos. Copernicus's idea was made against the background of the Ptolemaic system that had the earth as the center of the universe, while Copernicus proposed that it was rather that the Earth revolved around the Sun. The Heliocentric model, widely disseminated by Galileo (1564–1642) – both by his propagandist use of the vernacular Italian, but also evidenced by his use of the telescope – transformed European ideas about the nature of the Cosmos. But this shift in the Western conceptual scheme proved – rather as the Church had feared – not merely to be a reconceptualization of the cosmos, but also the human place within it. With the earth no longer at the center of the cosmos, the self-conception of human beings as existing in a universe specially created for us by God was deeply disturbed. This was a first blow was struck against the doctrine of human exceptionalism.

Charles Darwin's (1809–1892) theory of natural selection delivered a second conceptual shock for it indicated that human beings were not specially designed by God but by the same "blind watchmaker" processes of natural selection as the rest of nature. After Darwin, the view that humanity exists as separate to and outside the rest of nature was fundamentally challenged. The third discontinuity was inspired by Sigmund Freud's (1856–1939) distinction between the conscious and unconscious mind.[7] Freud claimed that his work "seeks to prove to the ego that it is not even master in its own house but must content itself with scanty information of what is going on unconsciously" behind the scenes.[8] These ideas directly confronted Descartes's notion that the conscious mind was diaphanous and open to itself, and raised the more worrying prospect that the deep motivations of our own behavior were hidden from us. Mazlish's case is that the information revolution and the construction of the computer is a fourth such conceptual shock.[9]

Another aspect of Freud's thesis was pointed out by developmental psychologist Jerome Bruner in his *Freud and the Image of Man* (Bruner 1956). Bruner noted that the "shocks to the ego", or discontinuities, described by Freud can also be viewed as establishing new *continuities*. The Copernican revolution, and its Newtonian extension, establishes that the heavens operate via the same laws as those that

[7] Freud humbly pointed to his own place in the history of ideas when he argued that his idea of the unconscious should be seen as a third discontinuity following the ideas of Darwin and Copernicus.

[8] This is cited in Mazlish (1993).

[9] The first three discontinuities were first described by Freud (1920). Luciano Floridi has recently developed a related thesis in his (2014) book *The Fourth Revolution: How the Infosphere is Shaping Human Reality* where he uses the nomenclature of "the information revolution" to label the fourth discontinuity. Floridi does not mention Mazlish's (1993) formulation although there are great similarities in the thinking as well as interesting differences between the two; we will discuss in the next section.

explain the movement of bodies on the earth. Darwin's ideas about the evolution of species showed that the same processes of natural selection that produced multifarious life across the earth also gave rise to the human species. While the Freudian idea of the unconscious, Bruner argues, showed that the same, biological laws of nature explained both the most savage episodes of human history and the heights of our civilization: a new sort of unified view of human nature.

If the third Freudian discontinuity revealed new vistas on our minds, the fourth discontinuity transforms the very notion of what a mind is. The fourth (dis)continuity can be variously described as mechanistic, computational or even informational. It is discontinuous in the sense that the human self-conception is radically reshaped from what was previously understood, and with this reshaping, a new shock to the ego is delivered. The fourth discontinuity is given to us by the information revolution, by the formalization of our understanding of computation and not least by the creation of computer technology itself. With the computer comes cognitive science as we understand it today and the idea that brains and indeed minds can be understood as encoders and transformers of information.[10] The fourth conceptual discontinuity recasts how we think of minds. Minds, thinking and all the processes of cognition are no longer conceived of as something immaterial but realized by specific mechanisms, especially the mechanisms of informational transformation and *computation*. With the computational revolution the construction of "thinking machines" becomes conceivable, and the idea and project of building artificial intelligence (AI) is revealed as a fundamental scientific and technological goal. Therefore, this discontinuity in the history of ideas, also reveals a deep underlying continuity, in this case, between the workings of human minds and the workings of the machines we create. The age of the fourth discontinuity is one where the once assumed to be special processes and inner realms of our cognitive life are increasingly seen as ones that can be modeled, simulated and even instantiated by computers and other human-built technologies. Viewed from the vantage point of continuity this conceptual revolution proposes that the same mechanisms that we use to build and explain "intelligent" machines also explain our own cognitive processes. And with this conception, the human self-image has once again been fundamentally altered.

[10] An idea resisted by many including Gibson 1979; Tallis 2004; and Varela et al. 1991.

1.2 The Information Age and the Computational Conception of Mind

Although the computer may be considered just another in the long list of technologies that human beings have used as metaphors to reframe their self-conceptions,[11] there is a difference. With the project of AI, two factors distinguish the importance and radical nature of the conceptual discontinuity that computers bring with them. The first is that the computer revolution does not just give us a new model of mind, but an understanding of the mechanisms that might allow us to build independently intelligent systems. The second is that the computational/information technology revolution confronts us with an ever-increasing range of "smart" systems that perform tasks that were once taken to be the sole province of the human mind. AI therefore unifies in one research program not just a wholesale reframing of what human minds are, but, by producing intelligent or "smart" systems that can do autonomously much cognitive work previously the province of the human intelligence, challenges the special nature of the human mind.

When did the first inklings of such an idea begin? Thomas Hobbes (1588–1679) significantly anticipates the idea that mental activity might be a form of computation already in Descartes's times. He writes about "reckoning", today we would say compation, that: "For 'reason' in this sense is nothing but 'reckoning,' that is adding and subtracting, of the consequences of general names agreed upon for the 'marking' and 'signifying' of our thoughts" (Boden 2006, p. 79). Hobbes' ideas may have been triggered by Blaise Pascal's creation, in around 1642–44, of a calculating device – today known as the 'Pascalina' – capable of four arithmetical operations. If so the idea that building of artefacts preceds theory has a pleasing early exempla.[12]

Yet, according to Margaret Boden, one of the foremost scholars of 'mind as machine', it is only in comparatively recent times that we have thought systematically of the workings of our minds in mechanist terms.[13] Boden argues that the notion is "more securely dated to the time of the second world war" (Boden 2006, p. 52). Specifically, Alan Turing laid the theoretical groundwork for the creation of the first digital computers (Turing 1937) and then considered that they might actually be used to model human intelligence (Turing 1950b). It was the development of

[11] The early modern period does give us a few well-known examples of technologies as metaphors for parts of the human body, such as the hydraulic metaphor used in the time of Descartes to illustrate the ways that bodies were supposed to move, to the mills of Leibniz's thought experiments. Later in the nineteenth century, telegraph connections were sometimes used as model of the interconnection of brains.

[12] Luciano Floridi (2014, p. 91) gives an interesting reconstruction of how Hobbes ideas from his 1651 treatise *Leviathan* may have been triggered by the creation and publicity of the of the Pascalina, the creation of which was influential throughout Europe.

[13] Boden observes in *Mind as Machine: A History of Cognitive Science* that the idea of "'Machine as Man' is an ancient idea, and a technical practice. Ingenious android machines whose movements resembled human behaviour, albeit in highly limited ways, were already built 2500 years. 'Man as Machine' is much more recent." (Boden 2006, p. 51).

the digital computer accelerated by the war effort which created a number of ideas around stored programs, the encoding of information, a general purpose computer and ultimately the thought that intelligence itself might be computational that laid the real theoretical foundations for the fourth discontinuity. Once it was possible to conceive of intelligent processes as algorithmic, it was only a short step to the notion that minds are computational. The theoretical possibility gave rise to the practical endeavor of building smart systems which is now reshaping the human world. With it we undergo a major conceptual shift as the mind-body problem is reconfigured into the mind-technology problem.

The first stage of the mind-technology problem is primarily conceptual, generating a new set of theoretical problems for philosophers and scientists. It is bound up with artificial intelligence as a project in research laboratories and the subject of speculation for philosophers. The second stage of the mind-technology problem, as we shall go on to describe, is established when we interact on an everyday basis with technological constructs that might actually be considered independently intelligent. The second stage gets underway as AI artefacts (smart technologies) start to become part of everyday life.

The first stage of the mind-technology problem becomes apparent when we not only conceive of our minds in terms of artefacts and mechanisms, but when we design, build and realize systems that are able to reproduce at least some of the mechanisms in functioning systems. The realization of AI then is central to this moment in the mind-technology problem. When we see mechanized systems able to carry out complex tasks autonomously that previously would have been seen as the exclusive province of human thought, a central domain of human uniqueness is challenged. Central moments in this development included building AIs that beat world champions in first Chess and then Go. IBM's Deep Blue and Deep Minds AlphaGo (respectively) were not just monumental technical achievements but existential challenges to the exceptionalism of human reason.

A second stage of the mind technology problem is less a conceptual challenge, than a practical challenge of how to live with our creation. Although AI applications have been increasingly permeating society over the last 30 or 40 years, it is perhaps really only in the last decade that "smart technologies" have become everyday interactants as parts of the daily lives a sizeable proportion of humanity. They have long been regulating our lifes in the background. As many of us now interact with "virtual assistants" such as Amazon Alexa or Apple Siri, they leap into the foreground, and a new form of existential challenge arises.[14]

The mind-technology problem resets the basic structural configuration of our understanding of mind. It can be regarded as the successor constellation to the mind-body problem, making new default assumptions about mind and world. Whereas the mind-body problem tacitly assumes that the definition of mental

[14] Hans Moravec (1988) calls these creations *Mind Children* and however we choose to treat them we cannot now doubt their centrality to our lives. It is because our further co-habitation with our creations is a central problematic of our age that Mary Shelley's *Frankenstein; or, The Modern Prometheus* (Shelley 2018 [1818]) remains the prescient touchstone text of our epoch.

processes is unproblematic and locates the basic difficulty in how our (ethereal) mental processes causally interact with matter, the mind-technology problem assumes that mental processes are material and that their definition is seriously problematic. Some problems such as the problem of mind-body interaction dissolve. If the mind is material, there is no mind-body problem as such, or at least no problem of interaction.

The mind-technology problem starts with the assumption that whatever minds and mental processes are, they are not a different type of stuff. The working hypothesis is that minds and cognition can be understood by understanding *mechanism*. It is those aspects of mind, i.e., intentionality, consciousness, agency etc. that may be argued not to be understood by mechanism that generate its characteristic range of problems. The problem can be resolved into three main theoretical components and a further practical problem.

1. What mechanisms are distinctive of minds and what if anything makes human minds and mental processes special?
2. How do Minds emerge from matter and material processes? Especially what kinds of mechanisms account for the distinctive cognitive abilities of minds?
3. What (if anything) demarcates the sorts of intelligent processes that are parts of natural minds from those that are instantiated in the artefacts we create?

More practically the mind-technology problem grows out of the ability to design and build actual artefactual systems that can exhibit intelligent behaviors, that is, it is arises from our confrontation with AI and (as they are increasingly known) Smart Devices. As we build such systems a fourth question arises, namely:

4. How are we to co-exist and live with the apparently intelligent systems we create? What should we hope and expect from them and how can we shape their future development?

We will now go on to look at several aspects of the mind-technology problem in a little more detail before discussing the contributions of the articles in this book.

1.2.1 AI and the Reconceptualization of Mind

If Alan Turing's work on computable numbers (Turing 1937) heralded the beginning of the information age, then his paper *Computing Machinery and Intelligence* (Turing 1950a) changed forever the way we conceive of the mind. Still, this vision took several decades to percolate through the world of ideas to the point where the explicitly computational program of cognitive science could be launched (Boden 1977, 2006; Gardner 1985). The computational model of mind (Fodor 1975a, b; Newell and Simon 1972) promised not just a metaphor of mind but – in the incarnation of AI – an approach to modelling and ultimately synthesizing the mechanics of human thought.

The founding moment of Artificial Intelligence as an explicit program of research is often dated to the 1956 Dartmouth Summer Research Project on AI organized by John McCarthy in Hanover, New Hampshire. The six to eight weeks of the event saw the attendance of a number of those who would later become luminaries of AI, including, McCarthy himself, Marvin Minsky, John Holland, Claude Shannon, Oliver Selfridge, Ross Ashby, Allen Newell and Herbert Simon.[15] The original funding proposal stated that "The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" (Moor 2006). This bold proposal was matched with an ambitious timeframe which suggested "We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer." (Moor 2006).

For the early years at least, the results of the new AI research program did not much live up to the boldness of the vision; and indeed, the over-zealous timeframe for delivery was to prove a perennial affliction. However, the meeting did manage to formulate many of the research directions and suggested many of the problems which would both drive and haunt AI research for the next half century. Despite significant heterogeneity of approaches in the initial meeting, much actual AI research would settle down to exploring the symbol system paradigm for the next several decades (See Boden 2006, Chapter 10) by which intelligent action was interpreted as the explicit syntactic manipulation of tokens that, in some way or other, represented a problem domain. Early successes included Arthur Samuel's (1959) checkers (draughts) playing program that received significant popular coverage for beating one of America's best players at the time. Board games were in many ways ideal system for the AI of the time as they involved problems spaces that could be readily represented by symbolic means. Despite some impressive successes in building systems to reason about such pre-specified 'microworlds', one limitation of the research was that this work could only rarely be 'scaled up' to deal with the dynamic, changing, open-ended and often ill-specified nature of many 'real world problems'. (Brooks 1990; Dennett 1984; Dreyfus 1972). Arguably, AI still faces these sorts of problems even though it has more recently supplemented the symbol system paradigm with a variety of sub-symbolic means of modelling cognition.

Another problem was that its successes were often over-interpreted or led the public to believe that AI systems had much deeper intellectual powers than they had. One famous program, Joseph Weizenbaum's Eliza (created in 1964–1966), was presented as a Rogerian psychotherapist that asked questions through a computer screen, beginning with the friendly question "Is something troubling you?" Participants would type their problems into a computer terminal and Eliza would

[15] The best and most authoritative account of the intellectual ferment that gave rise to Artificial Intelligence can be found in Margaret Boden's 2006 two volume set *Mind As Machine: A History of Cognitive Science*. Boden discusses "three inspirational interdisciplinary meetings", including the 1956 Dartmouth meeting, but also the IEEE 3-day symposium convened at MIT in mid-September and a later meeting held in 1958 in London as times and places where the central ideas of AI really germinated (see Boden 2006, Chapter 6.iii).

respond by asking further plausible sounding canned questions, sometimes eliciting rather complex responses from its human interactants (Weizenbaum 1976). 'Conversing' with Eliza was reported as a valuable experience by some of its users who often expressed the desire to go on using the system. However, Eliza was far from intelligent. Her programming was merely taking advantage of some fairly shallow syntactic processing. This, alongside the tendency of at least some human users to project much more intelligence onto the system than was actually embodied in its programming, gave the appearance that Eliza was asking penetrating therapeutic questions, whereas the system embodied no knowledge at all. It could merely to recognize and respond to some fairly crude and superficial features of human natural language.

It would be wrong to present all of this early work in AI as simulacra and smoke and mirrors. By the 1970s an extension of symbolic processing techniques, led to the *Expert System* paradigm which employed 'knowledge engineers' to interview domain-experts in order to symbolically encode and process their knowledge and inference patterns in order that their reasoning could be replicated by digital computers. This 'expert knowledge' could then be used to simulate expert performance. This approach had some very notable successes. MYCIN, for example, was an early but highly sophisticated expert system developed by Edward Shortliffe that was developed in order to diagnose infectious diseases. It was able to interact with professionals by asking questions to illicit information and express diagnosis including levels of doubt. It was found to outperform junior doctors in its diagnostic abilities and was widely used as a medical support system by practitioners.[16]

If computationalism was the bedrock and theoretical framework for much of the then burgeoning cognitive science, then the closely related discipline of AI provided much of the theoretical core of the new interdiscipline. Early AI focusing on the symbolic systems paradigm was the core of much cognitive modelling in the early days and also provided a mechanistic grounding for conceptualizing cognition. In addition, neural modelling – especially Rosenblatt's early work on the perceptron algorithm – attempted to model intelligence at the level of the brain. This program was killed off for more than a decade by Minsky and Papert's (1969) critique *Perceptrons: An essay in computational geometry.* Nevertheless, neural simulation was to be reborn through the development of the backpropagation algorithm and parallel distributed processing approach to cognition (Rumelhart and McClelland 1986a, b). From the early days in cognitive science, AI researchers were able to build many penetrating models of cognitive processes offering powerful models of (usually) human cogntion and including creative processes that were often held to be beyond the purview of mere machines (Boden 1990). This work paved the way for today's deep learning algorithms (Bengio 2009) that supply much of the algorithmic bases behind Google's search monopoly in so many areas of digital life. Rodney Brook's experiments in situated robotics (Brooks 1991a, b)

[16] For further details and excellent discussion of early work in artificial intelligence including its many successes and its limitations see Boden 2006, Chapter 10: When GOFAI was NEWFAI.

provided the anti-thesis to the main thesis of computational approaches to AI emphasizing situated-action and embodiment. Today, rather than opposition, they seem to be part of a new embodied/predictive synthesis in the explanation of the mind (Clark 2015a).[17]

In the midst of these early adventures in AI, two intertwined and overlapping approaches or tendencies in research emerged that would shape much thinking and launch a thousand controversies. These were the engineering approach to AI and the scientific approach to AI. According to the engineering approach to AI, it was relatively unimportant how a given intelligent process was realized. The goal of AI was simply to create useful systems that could do work would require intelligence if carried out by human beings (e.g., MYCIN might be such a system, albeit one closely modelled in a high-level description of human knowledge and reasoning). The way in which a system carried out its task was not important so long as it was effective and AI systems did not need to model human the actual mechanisms of human thinking in any strong sense.[18] Scientific AI by contrast set out to understand and, if possible, reproduce some of the actual mechanisms involved in human cognition. This enterprise, aimed at understanding actual other minds, was closely allied to cognitive science and might try to model and simulate cognitive processes at many different levels of abstraction. In some cases, the idea was that the best way to understand minds was to actually try to build them (Dennett 1978), including – later on – robotic systems (Brooks et al. 1999).

The scientific incarnation of AI played a pivotal role in establishing a different tradition and was considered a central part of the new minted interdiscipline of cognitive science. The famous nomenclature of strong and weak AI was only introduced somewhat later by John Searle in his attack on the idea that symbol processing systems could literally think, have subjectivity, be conscious or be considered to be a kind of mind (Searle 1980). The idea of strong AI should not be strictly identified with scientific AI. It is possible, and indeed much AI work does seek to model human or animal minds and varying degrees of abstraction without making any assertion that those systems are or could be actual minds. Nevertheless, the two faces of AI: the "weaker" engineering program, and the "stronger" scientific program continue to interlock up until the present day.[19] As we shall see however both

[17] Questions about whether predictive processing constitutes the *real* mechanisms of mind go far beyond the scope of this introduction. The interested reader is referred to Clark (2015b), Hohwy (2013) and the Chapter by Paul Smart (this volume).

[18] This weak AI can now be seen very much in the tradition of Pascal's calculating machine which was supposedly designed in order to ease the burden of the laborious calculations that Pascal's father needed to perform as part of his work as a supervisor of taxes.

[19] Today the term Artificial General Intelligence (AGI) is sometimes used in order to describe artificial systems which have human level intelligence (e.g., Goertzel and Pennachin 2007). It is important to note however that even an AGI that could replicate all the factors of human intelligence would not necessarily be a strong AI. It is conceptually possible to build an AGI that could match or even outperform human beings in any or every particular domain but still not be subjective in Searle's sense. Whether this is because, as Searle argues, symbol processing systems are just not the right sort of mechanistic systems to be subjective is an open question. Recent work in

Strong and Weak AI play a significant role in setting up the Mind-Technology Problem and how we construe it. Yet, even weak AI can play a role in both the way we conceptualize cognition and minds, but also importantly in reconstituting the nature of human cognition. It may also be that our extensive interactions with weak AI are already having profound effects on human cognition (See discussion in Sect. 1.2.3).

Alongside the theoretical conception of strong AI was a new metaphysical view of the mind that promised a novel approach to the mind-body problem; or, more accurately a fundamental reconfiguration of the constellation of problems therein encompassed. Functionalism became the new standard philosophical doctrine, superseding – but also integrating – aspects of both the mind brain identity theory and behaviorism (Kim 2006). While it is surely possible to formulate the basic ideas of functionalism without making reference – at least very explicitly – to computational states (e.g., David Malet Armstrong 1983),[20] ideas of computers and computationalism provides much of the conceptual apparatus and motivation for functionalist thinking. Functionalism heavily lent upon the notion of computation to make good the claim that a mental state could be realized in a number of different potential implementations (or realizers) (Putnam 1980), just as software can be realized on a variety of different hardware.[21] Thus, functionalism can be seen as the distinctive philosophical position of this new informational period with its origin deeply tied to the theoretical and practical developments of computer technology. Further, functionalism helped to articulate fundamental problems in new ways. Whereas the then current mind/brain identity theory seemed to push us toward a too close identification of mental states with brain states, functionalism made it possible to articulate and imagine mental states as being realized by an increasingly exotic set of realization bases, from Martians in pain to computational systems that implement minds. Functionalism undoubtedly left many problems unsolved, not least the question of how to account for consciousness (Armstrong 1980; Chalmers 1995; Searle 1980), but it also laid the theoretical foundation of the new constellation around the nature of mind: the mind-technology problem.

Some of the most important implications of functionalism were not noticed until much later. Functionalism leaves open significant questions about the boundaries of mind. The idea of the extended mind (Clark and Chalmers 1998) – as we shall further discuss in Sect. 1.2.3 – suggests that the causal functional profile of the mind needs not to be implemented by the brain alone. Once functionalism makes it possible for us to conceive of how the mind might be multiply-realized (Putnam 1967),

machine consciousness (e.g., Clowes et al. 2007; Holland 2003) seeks, among other things, to attempt to understand if non-organic mechanistic systems – predominantly computational systems – could ever be subjective or put another way, conscious.

[20] Armstrong's much collected and influential essay from the book "The Causal Theory of Mind" (Armstrong 1980) formulates a causal analysis of mind in functionalist terms without mentioning computers; although he does imply that perceptual states in the brain are informational states.

[21] The idea that the mind is literally software for the brain remains controversial and has recently come under sustained attack (Piccinini 2010).

it is only a small theoretical step to conceive of how the realization basis of the mind might not just be the brain – or even the body – but spread out from this cognitive core to the super-dermal world beyond. Also, recently, the computational underpinnings of functionalism have been more deeply probed and these have revealed a series of novel problems over how exactly we think of the personal nature of mind. The notion of human personal identity for example may be difficult to make sense of in terms of programs and computational concepts.[22] Several of the chapters in this book explore how these notions of extended mind and personal identity interact, and explore how adequately the computationalist framework may be to support them.

It is important to note that not all cognitive scientists agree with, or continue to employ, the standard computational model of mind (Fodor 1975a, b; Newell and Simon 1972), certainly not as it was framed in the early days (Milkowski 2013; Schneider 2011). Indeed, many theoretical programs in cognitive science, such as many of the various forms of enactivism (Varela et al. 1991), are explicitly conceived of in opposition to this computational model. Some also explicitly reject the notion of informational and computational theories of mind (Tallis 2004) as well as the idea that human minds can be modelled – much less instantiated – by computers (Dreyfus 1972; Searle 1980). Yet, computational ideas of mind, and specifically its promise to allow us to model and even synthesize cognitive processes, has fundamentally reposed our understanding of what minds and mental processes might be, and what kind of systems might be considered to instantiate them.

The mind-technology problem emerges from a certain sort of – perhaps unstable – resolution of the mind-body problem in terms of functionalism and the computational model of mind. But not only does it offer a novel framework for thinking about minds, but it poses a host of distinctive questions. If our minds can be implemented by machines – in particular by computational systems – what, if anything, is the difference between our minds and theirs? Is it merely that we have been engineered by natural selection and not human engineers or scientists? That we are biological and that they, up until now, are largely implemented in silica? What properties of minds like ours can be implemented by machines – computational or otherwise? Can we, in reality, hold onto the theoretical distinction between strong and weak AI? Or is the boundary between minded systems and those that do merely "smart" computation, blurry, and indistinct? In this context, questions such as exactly how we should work through the software/hardware duality of the mind, or whether consciousness can ever receive a functional or mechanist explanation (Chalmers 1995; Dennett 1996a) are core problems for the new philosophical outlook.

[22] See the Schneider and Corabi paper in this volume, but also Schneider's book *Artificial You: AI and the Future of Your Mind* (Schneider 2019) for an extended and highly illuminating discussion. The burden of her recent book – and several essays in this one – is that the idea of the software metaphor of mind creates multiple problems, not least when we consider the notion of personal identity (see the papers by Schneider and Corabi, and Piccinini this volume).

1.2.2 The Information/Computation Revolution as Cognitive Transformation

According to Mazlish there are two separate theses that compose the fourth discontinuity and, we can add, help us pose the mind-technology problem with more bite. The first, that we have now amply discussed, is that human technologies, especially in the form of computer technology, not only serve as a model for conceptualizing minds, but can be re-used to explain the workings of our own minds. As Mazlish wrote 'we are coming to realize that humans and the machines they create are continuous and the very same conceptual schemes that help explain the workings of the brain also explain the working of a "thinking machine".' (Mazlish 1993, p. 4). This, it has to be said controversial "realization", lies at the heart of the mind-technology problem and establishes the new constellation of philosophical problems in which we are currently enmeshed. The second, and if anything, even more controversial aspect, is the realization that human beings, our concepts, our cognitive abilities and even the sorts of minds we possess have been fashioned through this process of making, that is, through the deep history of our interactions with artefacts and technology. For Mazlish, we cannot adequately conceive of human abilities and human cognition without factoring our "nature" as makers of artefacts and technology.

The human mind is thus conceived of, not just as a straightforward product of natural selection but is itself produced through the deep history of the construction of artefacts. From the creation of the first Acheulian hand-axe (Mithen 1996), through a process of increasing refinement and diversity, first slowly and then with rapidly accelerating pace, we have constructed a vast variety of artefacts and technologies, that have time and again changed and reforged the material culture on which we come to depend. With this reforgement, we have reshaped the ecological niches we inhabit, opening new behavioral possibilities for ourselves and making the possible the development of new skillful practices and forms of cognition (Malafouris 2013; Menary 2014; Sterelny 2011). But through the same process, and through our intimate reliance on the artefacts we create, we have progressively refined and variegated our cognitive abilities and cognitive potential. Our creation of artefacts to better shape the world to our own purposes has reciprocally transformed the nature of the human cognitive profile. We have recreated ourselves in the image of our tools. We are thus both natural beings and also in a certain sense, self-constructed, i.e., we are not just Homo Faber, man the maker, but human beings the self makers.[23]

The crucial question for us however is what this means for the nature of our minds in the computational age. Does the artefactual world we are building in the twenty-first century, and the digital information processing technologies that are increasingly embedded throughout all aspects of our material culture, alter the nature of our cognition and the nature of our minds? According to one influential

[23] For a recent exploration of this theme see Ihde and Malafouris (2019). The idea however has a long history (Vygotsky and Luria 1994).

strand of contemporary thought, while our technology undergoes dramatic changes, our minds and crucially the information processing profile of our brains remain substantially the same. Evolutionary psychology in its strongest form holds that the brain is like a swiss army knife where human cognition is defined by a set of domain specific cognitive apparatus designed by evolution to ensure our survival on the African Savannah (Barkow et al. 1992). On this view, the development of technology does not afford new cognitive potential so much as offers a new landscape for which brains and thus minds are ill-adapted. We are "Junk Food Monkeys" (Sapolsky 1997), forever doomed to inhabit a bleak technological landscape with ill-adapted brains.

The alternative *Homo Faber* view, which holds that the unique cognitive abilities of human beings are dependent upon the history and pre-history of our fashioning of artefacts, is, once again, gaining ground (Ihde and Malafouris 2019). If this is correct–as we have just described–the human mind is not strictly a product of natural selection, but a product of the new developmental and evolutionary pathways we have opened for ourselves through the fashioning of tools and artefacts (Malafouris 2010b, 2016). The question then becomes, how might the creation of information technologies, and the new behavioural and cognitive possibilities they afford, allows us, or, more pessimistically, unconsciously lead us, into transforming our cognitive capabilities as we interact with, and come to rely upon, a new order of ever-present ambient computational technologies.

The self-becoming of homo sapiens can be traced through signal moments in the history of the production and deployment of artefacts, from the slow development of the Paleolithic hand axe and its possible role in the development of human fluidity of thought (Mithen 1996), to the making of bronze age tokens that drove our mathematical capabilities (Malafouris 2010a). Even the development of distinctly human agency may be closely tied to the creation of tools which allow us to track our projects and more directly self-shape (Clowes 2019; Knappett and Malafouris 2008; Vygotsky 1962). Human beings, through the making of tools, open up new developmental trajectories and new cognitive possibilities for themselves and for future generations. Once humans start to have a complex tool-using culture the ability to refashion ourselves and our minds–albeit in the first instance largely unconsciously–becomes extensive. One benefit of explicitly formulating the mind-technology question, as we have done here is that it opens the possibility of more seriously and consciously intervening in the design of technologies that will shape our minds, and the minds of future humans.

Understanding the creation and use of artefacts thus becomes an inseparable dimension not just for understanding the genesis and nature of the human mind but also for understanding and perhaps shaping its future. If becoming human is thus closely tied to the history of our use of artefacts, might the future of the human mind – or possibly the post-human mind – similarly depend on the nature of the technologies we now rapidly deploying throughout our civilization?

The claim that human nature cannot be understood separately from our technological and artefactual culture is therefore a central aspect of the mind-technology problem. At one level, this can be partly understood as a form of niche-construction

and has some analogues among the shaping of habitats and reciprocal dependence on those shapings that we find anticipated among the habitat shaping of other animals (Laland et al. 2000; Menary 2014; Sterelny 2011). However the human artefactual world is also unprecedented not just in its variety and our wide use of artefacts to create more artefacts, but also in the way that we employ them to expand our cognitive capabilities (Gregory 1981; Malafouris 2013; Vygotsky 1978). Indeed, many of our technologies might be better conceived of as *mind tools* whose primary job is not to help us shape the world but our own cognition (Dennett 1996b; Gregory 1981). Moreover, according to recent theoretical developments, our artefacts are not just parts of our environments but – according to the extended mind view – also part realize our mental states (Clark and Chalmers 1998).

The Extended Mind Thesis (EMT) is a radically anti-Cartesian view that holds that human minds can sometimes come to rely on external artefacts through such dense patterns of interaction that those artefacts can come to count as part of the system that realizes a human mind. The thesis has become a central philosophical pivot of our intellectual moment and allows us a new way to articulate how it is that minds and technology are intwined. It also invites a promising resolution to the mind-technology problem. First Wave approaches to the EMT emphasized the *parity principle* according to which an artefact or system that provided the same functional profile as a part of the brain, and fulfilled the famous trust and glue conditions, could be considered a part of an individual's mind.[24] Second wave approaches to EMT instead foreground the *complementarity* of the artefacts. The way that they can provide different and novel cognitive functions for the mind that can be very different from our native or non-enhanced cognitive profile. Second Wave approaches emphasize how artefacts can bring properties that complement our native cognitive profile (Sutton 2010). The cognitive history of human beings can thus be viewed as process of the innovation and accretion of new cognitive functions through our deep and interpenetrative relationship with technology (Ihde 1990). This is an essential part of what has made us humans, and thus, we are, it is claimed, *Natural Born Cyborgs* (Clark 2003); that is, beings who get their particular species nature from the role that technology plays in our minds. From this perspective, the history of our tool use, even prior to the advent of AI, is literally a process of mind design.

Not everyone agrees with this perspective on technology. According to Luciano Floridi's (2014) book *The Fourth Revolution*[25] the informational revolution is fundamentally transforming the landscape that we and our minds inhabit, and indeed transforming human reality in the process, even if our minds and cognitive abilities are left more or less the same. For Floridi, we are implicitly coming to think of ourselves as *Inforgs,* or "informationally embodied organisms". While he rejects the cyborg vision of humanity, he thinks we nevertheless need to take account of the real changes that have taken place in the human environment as we interact and

[24] Further discussion of EMT can be found throughout book especially in Chapter XYZ.

[25] Its full title is *The Fourth Revolution: How the Infosphere is Reshaping Human Reality.*

cohabit with an increasing range of ever more autonomous non-human information processing systems. His thought is that the human beings are increasingly viewed as just one information processor among many, even if, as he argues, human and computational intelligence are really radically discontinuous. He writes: "We have begun to understand ourselves as Inforgs not through some biotechnological transformations in our bodies, but, more seriously and realistically, through the radical transformation of our environment and the agents operating within it." (Floridi 2014, p. 96). Our sense of self has thus undergone a radical shift, if not so much the nature of human cognition and the mind.

The concept of inforg is used to express the enforced parity to the ways human beings and the other information processing systems interact in the new terrain of "onlife". Yet, for Floridi, computers are just syntactic processors that lack the intrinsic meaningfulness and original intentionality of human minds.[26] Our conceptions of mind are thus undergoing profound change even if the cognitive profile of our minds is more or less as it was before. (Floridi 2014, 2015). He thus offers an explanation of why we tend to conceive of ourselves in technological ways, yet he rejects the cyborg/extended view of mind according to which human cognitive dependence upon technology is much stronger.

Increasingly there is less sense in viewing the informational world as a separate "virtual" world, but through innovations such as smart phones, the internet and wearable interfaces to the web our devices are increasingly becoming an interface to a second world of information. Moreover, this informational world is rapidly being engineered so that our computational systems can interact ever-more autonomously with each other through it. They are increasingly the natives of this realm. Floridi describes the way there is increasingly less distinction between the online world and "real life" as progressively changing the way that human beings view both themselves and the nature of reality. Human life has progressively become an "*onlife*" constantly accompanied by a digital shadow of existence of the informational world which increasingly interpenetrates with our world of lived experience (Floridi 2015).[27] In onlife, human and artificial inforgs interact as apparent equals despite their radical difference.

The idea that we now live in a computational or informational age remains controversial but is not new. Versions of this idea have been around for at least the last 50 years (Castells 1996; Toffler 1980). Floridi's particular spin on this picture rests on a historical analysis and periodization of human interaction with technology and how this changes our conceptualization of ourselves and our place in the world. Floridi separates human development into three phases or ages: prehistory, history

[26] In this, the foundations of Floridi's view clearly echo John Searle's (1980) position that no amount computational syntactic processing can ever add up to the inherent meaningfulness and subjectivity of real minds.

[27] In some respects, the notion of human reality is a little vague and seems to be used in a variety of different ways by Floridi, but a central meaning is how we now are coming to thinking of the different between "real life" and the virtual and online world. It is Floridi's claim that this distinction is increasingly breaking down and we are already starting to live in a blurred "onlife" reality.

and hyper-history, the latter of which roughly maps onto what is now often called the information age. For Floridi the information age is marked about our increasingly dense interaction with the information processing systems that now envelop our lives. We will briefly retell this story here, but drawing out–somewhat contrary to Floridi's own telling–some of the cognitive adaptations, indeed transformations, that take place with each epoch of technological development.

Pre-history was not without technological development, which had–on at least some analyses–momentous cognitive import. Steve Mithen (1996) for instance argues that the several phases of development of the stone axe fundamentally change the representational abilities of the human mind including the development of cross-module cognitive fluidity. If we take such cognitive fluidity to be a special feature of human cognition, and if Mithen is right, one of the central attributes of the human mind arose in the very particular context of artefact development.

Around 12,000 years ago humans developed agriculture. This technological revolution changed us from being primarily nomadic hunter-gatherers to a sedentary species. Through manipulation of plant species and animal husbandry human beings were able to change their means of subsistence and life. The transition from hunter-gatherer to farmer makes possible not just new ways of subsistence but new ways of being human. Taking up the sedentary life may have many important cognitive advantages, but it is likely that settled communities created enhanced conditions for passing on knowledge to future generations and perhaps the origins of some of the forms of cognitive apprenticeship which allowed the development of richer material – and other forms – of culture (Sterelny 2011).

In Floridi's telling it took more or less 6000 years for agricultural technology to produce its fruit in the shape of a new technological revolution. Around 4000 BCE, some Bronze Age societies begin to develop the largely overlapping technological inventions of writing, bureaucracy and cities. Cities again transformed human beings and their society. Amongst other aspects of this transformation was the specialization of labor resulting in the burgeoning of dozens of different crafts along with new tools and attendant techniques. We can add that these civilizational developments go alongside a cognitive transformation especially that associated with the development of writing and reading Learning to read has long been argued to promote a cognitive transformation (Luria 1976; Ong 1982) albeit to much controversy. In many ways, this controversy was laid to rest as the neural and cognitive implications of learning to read have now been demonstrated in impressive detail (Castro-Caldas et al. 1998; Dehaene 2009).

In the late twentieth and early twenty-first centuries, and roughly 6000 years after our first development of writing – that is, our first major information communication technology (ICT) – we are once again undergoing another such transformation which is the informational revolution he calls *hyperhistory.* For Floridi, hyperhistorical societies are those where the recording, movement and control of information takes up a larger part of the economy than manufacturing or agriculture. There is some impressive empirical data to back up this claim, not least the massive increasing in the production of information. The consequent need or desire to try to analyze all this new information, i.e., the problem of *big data*, is another

characteristic one of our age. In the G7 countries most of the GDP is no longer made up by the production of material goods, but from the service economy and what is essentially the production, control and administration of information. And just as more economic activity is now generated by moving around bits than material goods, the digital realm has become ever more deeply embedded in the second-by-second existences of many – perhaps the majority – of citizens of the economically most advanced nations.[28] Floridi writes that hyperhistorical societies are those in which "ICTs and their data-processing capabilities are not just important but essential conditions for the maintenance and any further development of societal welfare, personal well-being, and overall flourishing." (Floridi 2014, p. 4).

Whether we are really entering such a dramatically new phase of human development is difficult to derive from any source of empirical evidence and tricky to precisely characterize. Human beings of course still live off the back of the agricultural revolution and just because much of the world's industrial production has moved to China does not mean we are no longer living in an industrial civilization. Even though the computer has undoubtedly penetrated an ever-increasing range of human activities, it still allows for the question about whether many of these activities are really different by nature and not just technical means. The world's first city, Ur already had a highly developed bureaucracy[29] and, as Floridi himself notes, was already a sort of informational society. Nevertheless, it is difficult to make the argument today that the computer is not a dominant feature of the most influential societies of the early twenty-first century and difficult to imagine non-catastrophic circumstances where this is not also a dominant feature of any near-future human societies. It is more difficult still to hold that co-existing with these smart technologies are not likely to have a profound effect on the nature of human cognition.

We might then accept Floridi's periodization of human civilizational development but reject his view that the nature of human beings is – so far – relatively closed to cognitive augmentation via ICTs. This outlook pushes us towards a more radical view both of our technologies and of our minds. This view is implicit in Clark's account of natural born cyborgs. But here we should make explicit the extra turn of the screw. If the nature of our minds is radically open to technological interaction as we here suggest, then the fact that we are now co-existing with an increasingly varied array of smart machines is likely to have profound effects on human cognition and the nature of our minds.

The latter part of the twentieth century and especially the early part of the twenty-first century signal a new phase of our encounter with AI where everyday life involves our deep involvement with a range of "smart technologies". Smart technologies are difficult to define but the central idea is that they embed in one form or

[28] Albeit see the discussion at the beginning of Sect. 1.3 on some of the cross-cultural diversity that there is in how humans have variously thought of AI.

[29] Ur is the original Mesopotamian city in modern day Iraq, once a coastal city near the mouth of the Euphrates and today part of dessert landscape. It appears to be the case that the building of cities, the invention of writing and complex bureaucracies are roughly coeval developments of the human species.

another some aspects of AI technology into their design. The importance of this encounter is implicit in the foregoing discussion, namely that technologies have long made a contribution to human intelligence in one form or another. Things, play a central role in making us smart, as Donald Norman (1993) has shown us. However, with AI the new and discontinuous factor with respect to the history of technology, is that smart technologies do not merely contribute toward, or mediate, human intelligence, but can be interpreted as being intelligent in their own right.

To understand ourselves, we need to take account of the radical change in the technological background as AIs, either weak, strong or in the varied space in between, become the ever-present background to all our cognitive processes. Human beings are engaged in deep interaction with new generations of smart technologies and the transformations this is likely to imply for the sorts of creatures we are has so far scarcely been considered. Let us formulate a new question based upon these considerations. What is it to be human in the time of smart technologies?

1.2.3 Being Human in an Age of Smart Artefacts

Questions about the relationship between artefacts and mind move into a radically new phase as our interactions with AI technologies becomes a factor of everyday life, especially as we come to depend on these technologies. The second phase of the mind-technology problem reflects the technological moment we are living through in which we encounter some form of smart technology on a daily basis. The everyday and ubiquitous nature of this technology is predicated on a new regime of Cloud-Technology, where so many of our artefacts present personalized, localized and often AI-enhanced functions in virtue of their connection to distant servers that provide services whenever they are needed. Many of our artefacts have become Cloud-Technology, i.e., technology connected to invisible systems somewhere in the cloud, and the cognitive lives of human beings has come to increasingly depend on this ever-present background (Clowes 2015). Deep theoretical questions now move from the seminar room to become the practical challenges that confront human beings everywhere, forcing us to reconsider the ways we organize our society and our individual lives. The challenge of how to live with AI systems becomes both a practical and a deeply ethical challenge. The central question of the 2nd phase of the Mind-Technology Problem becomes: *How are we to live in a world of ubiquitous smart artefacts?*

Forms of AI derived technology are becoming ever more pervasive in society from the invisible systems setting our credit ratings, to the AI embedded in systems helping us finding our way to a restaurant, or a date, to the virtual agents such as Amazon Alexa or Apple Siri that we consult on our mobile devices or the smart speakers that reside in our homes. AI in the form of smart technologies presents itself as a practical challenge as some of the more metaphysical questions may fall into the background. However, central questions such as what the status the sorts of intelligence we create should have, are never far from the surface. How far, and

under what circumstances we can rely, and should we rely, on such intelligences, whose ultimate status vis-à-vis our own is undecided, continues to be a nagging problem.

The status we accord to smart technologies is inseparable from the way we regard our own minds. One approach is to argue for a strong divide between the artificial intelligences we build and the natural organic intelligences of which we might take ourselves to be a paradigm. Floridi for instance argues that the sort of AI that we encounter in the form of smart artefacts is only 'Light AI' and should regarded as type different from our own. Light AI, it is said, relies on purely syntactic processing (Floridi 2014, p. 141) and that it is not really intelligent at all.[30] Floridi here is largely following Searle's distinction between Strong AI which would purportedly have human-like intelligence and possibly other cognitive attributes and Weak AI that is only a sort of syntactic processing.[31] On the interpretation favored by Floridi, AI technology now and, in at least the near future, is likely to be Light AI and therefore we should always keep in mind the differences between ourselves and the apparently smart systems with which we interact. The technology is in a sense the inheritor of the Eliza program, with the apparent intelligence these systems provide being largely a projection. While AI may accomplish useful tasks for us, it is primarily doing so by blind syntactic processes which, while they may be useful (smart), share little with human (intelligent) cognitive processes. Floridi argues that "The fact that in 2011 Watson—IBM's system capable of answering questions asked in natural language—won against its human opponents when playing Jeopardy! only shows that artefacts can be smart without being intelligent. Data miners do not need to be intelligent to be successful." (Floridi 2014, pp. 140–141). Whether a sharp distinction between artificial smart technologies, and real human intelligence can really be made to hold is a controversial question. As AI researchers continue to build and deploy technologies based on deep analyses of human cognition the line seems set to be ever more blurred.

Regardless of its status, AI of one sort or another is increasingly central to much of the informational traffic that regulates our lives. The appliances and gadgets we carry with us, often on our mobile "smart" phones connect directly to the internet and have become the constant accompaniment to our cognitive lives. With 4G and

[30] See Chapter 6, "Intelligence Inscribing the World" of Floridi (2014) for a treatment of how he sees human civilization adapting to the reality of cohabiting with light or weak AI. Light AI does not appear to be fully defined but functions in the discussion to pick out the sort of intelligence we find in smart systems to which no true intelligence should be attributed; (whatever real intelligence is).

[31] Floridi writes "The two souls of AI have been variously and not always consistently named. Sometimes the distinctions weak vs. strong AI, or good old-fashioned vs. new or nouvelle AI, have been used to capture the difference. I prefer to use the less loaded distinction between light vs. strong AI." (Floridi 2014, p. 141) While Floridi is right about the inconsistent naming this explanation risks confusing matters further. Strong and Weak AI were originally used to distinguish two approaches what AI was supposed to be doing, whether buidling systems that might really be subjective or minds, or merely doing a form of non-subjective processing. GOFAI and nouvelle AI were different approaches to how these different goals could be obtained (See e.g., Brooks 1990).

now 5G networks rapidly being deployed and as the internet of things is becoming an everyday reality (Smart et al. 2018), however we interpret the nature of the intelligence they embody, smart artefacts are becoming ever-present parts of our lives. As our "native" cognitive processes come to increasingly rely upon an environment densely populated with smart artefacts the character of human cognition may be undergoing profound changes. For many, this new situation will suggest the extended mind view as a natural framework to interpret how human cognitive may be changing in accommodate these new settings. Not everyone agrees. Floridi writes that "the view according to which devices, tools, and other environmental supports or props may be enrolled as proper parts of our 'extended minds' is outdated. It is still based on a Cartesian agent, stand-alone and fully in charge of the cognitive environment, which is controlling and using through its mental prostheses, from paper and pencil to a smartphone, from a diary to a tablet, from a knot in the handkerchief to a computer." (Floridi 2014, p. 95). But this misrepresents the basic idea. On the Extended Mind view the agent is not understood as "stand-alone" or its agency separated from the artefacts and systems on which it depends. Rather the agency of the cognitive system is understood to be distributed among the components both organic and technological (Clark 2006). The radical nature of the vision is that human agency is dependent upon, shaped by and may even partly incorporate parts of the artefactual world.

The idea of the Extended Mind which may have seemed like a distant and exotic theoretical possibility when the idea was first mooted in 1998,[32] now, in times of the ever-deepening reliance of many millions of people on smart gadgetry, has become something of a banal reality (Clowes 2015). In the age of the smartphone, pervasive computing and the everyday presence of AI systems such as Amazon Alexa and Apple Siri, it also becomes a central element of our Weltanschauung.[33] Today, the idea that ICTs can be part of our minds no longer seems so outlandish and may even be becoming part of folk psychology.

Perhaps the most telling examples involve how we seem to be rapidly reconceptualizing the nature of human memory through our dense interactions with E-Memory devices. E-Memory systems are digital electronic systems or devices which replace, extend or augment human biological memory.[34] Our constant access to technologies that can provide E-Memory functions may already be profoundly reshaping human organic memory. According to the so-called google effect, it is now reported that some users of internet systems may preferentially "remember" facts about how to access information with their favored ICT tools rather than remember the actual information itself (Sparrow et al. 2011; Wegner and Ward

[32] See for instance the discussion in Fodor (2009).

[33] This is one reason the original extended mind paper is the most cited philosophy paper of the last 20 years.

[34] E-Memory might also be defined as a heterogenous set of digital or electronic systems that provide similar or replacement functions that would otherwise be provided by human biological memory, see (Clowes 2013) for further discussion and the slightly problematic nature of these definitions.

2013). Some subjects go a step further reporting that information they can readily access from their smart phones constitutes their own knowledge. This fact is often reported as an epistemic error, but it raises difficult questions about what constitutes knowledge when our epistemic environment undergoes such profound changes (see Clowes 2017). If the internet connected smartphone now just constitutes part of the reliant and ever-present environment, might at least some of the information we access with these devices actually count as our own knowledge, even as part of our own personal memory?[35] Our intimate encounters with and increasing reliance upon these technologies may thus be rapidly changing how we conceive of human memory (Clowes 2017), what human memory is, and the role it plays in human attributes such as personal identity (Clowes 2013, 2020b; Heersmink 2016, 2020).[36]

In this way, the mind-technology problem becomes implicated in our folk-psychology, i.e., in how everyday folk conceive of what a mind is and what a mind does. Tad Zawidzki's notion of mind shaping is a useful theoretical perspective here (Zawidzki 2013). According to Zawidzki, folk-psychology, once thought of as the pre-scientific series of intuitions and interpretative mechanisms we use to make sense of the mind (Wilkes 1984) should also be thought of as a social mechanism that plays a central role in developmentally constituting human minds as such.[37] It is through the personal history of being interpreted as a mindful agent with believes and desires, and interacting with others that are so interpreted, that children come to conform more closely to norms of society. Folk psychological interpretations and narrative practice can be seen in playing a central role in constituting minds as such (Gallagher 2001; Hutto 2008; McGeer 2001).

What might be the consequences of these practices coming to accommodate our interactions with, for example, virtual assistants? Especially since some such systems are now being used by very young children. If the nature of our minds is so dependent upon human folk-psychological practices of interpretation, how might those practices change as we adjust to incorporating the likes of Siri and Alexa into our social lives? Some have worried that the human adjustment to a shared social space that includes AI generated interlocuters may fundamentally alter human socialization and the nature of our social interrelations (Turkle 2011). One implication of this second phase of the mind technology problem is that we may come to reimagine human nature itself through the prism of our interaction with artificial

[35] Some of the questions about the nature of the epistemic framework we can use to accommodate an increasingly diverse world, cognitive agents and their relations with, on the one hand, technologies and, on the other, social practices are treated in Gloria Andrada's paper in this volume.

[36] It is only in the time of cloud-technology that we could begin to seriously worry that our minds were leaking out to machines in the way Nicholas Carr articulated (Carr 2008, 2010). It is important to see that this is a residual of how we think about our minds in relation to the current form (cloud tech) and deep tendencies (e.g. Moore's Law, pervasive computing) of computer technology. These problems are unlikely to subside anytime soon and many of the intellectual tools we need to grasp have yet to be invented. The hope is that explicitly laying out some of the distinctive difficulties of our conceptual epoch in this volume we can move forward.

[37] For a predecessor view to Zawidzki see McGeer (2001). Also of relevance is Gallagher (2001).

systems. Some therefore worry that a sort of false identification with our technology may be the consequence and that it will undermine our humanity (Lanier 2010).

The question of how human beings should regard the increasing penetration of the internet and AI technologies into our lives has been one of the most controversial and polemical questions in recent times. On one analysis, our tendency to use the internet as the central source of intellectual reliance is depleting our minds and turning us into more shallow beings (Carr 2008, 2010; Greenfield 2015). The internet and ever growing raft of AI technologies with which we interact is undermining our relationships and social cognition (Turkle 2011), our memories (Sparrow et al. 2011) and individuality (Lanier 2010), and even human agency (Loh and Kanai 2015). According to these views our reliance upon smart technology should primarily be viewed as a danger to the human mind. A more nuanced picture acknowledges how technology has always been a central part of the human life-world, and in part confers our cognitive abilities. The internet and smart technologies are, on this view, seen as part constituting a new *cognitive ecology* upon which human beings can draw and which can provide new cognitive potentialities (Smart et al. 2017).

On one view, our increasing use of smart technologies can be seen as a sort of outsourcing by which our most important cognitive abilities are increasingly being carried out by machines (Gerken 2014). Another, not necessarily contradictory view suggests they may be being incorporated into our cognitive lives in ways that could add up to a new kind of human agency (Clowes 2019). If smart systems can become proper parts of our minds, especially if they are understood as Light AI in the way Floridi suggests we might view this as a less worrisome sort of interaction than the outsourcing account. Which picture better fits the realities of our dependencies on smart technologies is a nuanced and still much underexplored territory (Clowes 2020b).

A further alternative articulated by Paul Smart (this volume) is of AI increasingly using technology that is not only modelled on cognitive mechanisms but is mechanistically realized in a manner that makes use of the same sorts of cognitive mechanism. Technologies such as deep learning (Bengio 2009), which has structural and computational similarities to predictive processing systems, are thought by many to be the main mechanistic underpinning to human cognition (Clark 2015b; Hohwy 2013). Clearly the nature of the AI technologies with which we interact is of great importance for how we should consider the nature of "intelligence outsourcing", not least because, on some interpretations of current AI technology, it is itself already highly autonomous.

However, we come to resolve these problems, the boundaries between us and our machines, will be further breached in the coming years. Susan Schneider (2019) argues in her recent book that the technologies of cognitive augmentation via brain implants that directly connect to digital technologies are already much closer to changing the nature of human futures than we think. Elon Musk's neuralink is just one such technology. In her book, Schneider imagines a "Mind Design" salon where – in the near future – people will be able to enter to commission upgrades to their cognitive abilities. If the line of argument pursued in this introduction, and

extended mind theory is correct, we have already been participating in a "home-brew" version of digital cognitive augmentation for some time. Schneider is particularly interested in questions how these ideas of cognitive augmentation relate to questions of personal identity. If I upgrade my memory to super-human levels, will it still be me who comes out at the end of the upgrade process? Schneider's thought experiment over the mind-design salon raises crucial questions for what and how our mind and technology interact that deeply implicate personal identity and continuity. E-Memory technologies such as the smart phone may already provide adept and reliant users with a form of cognitive augmentation that can play a potent role in personal identity (Clowes 2013, 2017, 2020b; Heersmink 2016, 2017).

Some contemporary AI ethicists labels these sorts of scenarios as science fiction (Coeckelbergh 2020; Floridi 2020), and argue that much speculative current work on artificial intelligence may be distracting us from the serious problems of living with AI as it actually exists today. Ideas about, for instance super-intelligence (Bostrom 2014) and existential risk (Russell 2019) merely distract us from current realities and difficult ethical problems produced by the 'light AIs" of today. One response comes from Stuart Russel who wryly notes that many of the same authors who in a rather boosterish manner, laud the open-ended possibilities of AI at the same time dismiss the existential risk. He notes that 'Within the AI community, a kind of denialism is emerging, even going as far as denying the possibility of success in achieving the long-term goals of AI. It's as if a bus driver, with all of humanity as passengers, said, "Yes, I am driving as hard as I can towards a cliff, but trust me, we'll run out of gas before we get there!"' (Russell 2019, p. 8). A related trouble though is that, as we come to increasingly rely upon often opaque smart systems, often built by private sector companies that may be reluctant to fully release the algorithmic details of their systems, it is becoming difficult to know where the science fiction ends, and the technological reality begins. Algorithms using systems like GPT-3 are rapidly redefining what is possible with today's technology. Only the very brave will hazard the limits of these technologies today (Benzon 2020; Floridi and Chiriatti 2020a). The real existential risk may be to not recognise the radical existental implications of technologies we are already living with.

The mind-technology problem exists wherever and whenever it is that our minds stop, and artefacts begin and inhabits the increasingly over-populated grey area where we are no longer sure which is which. This difficult grey area will take up much of the discussion in the rest of this book.

1.3 Reconceiving the Mind in a Time of Smart Technologies

Even though the actual status of current and future AI is much disputed, our present encounter with it is rapidly reframing how we think of our own minds. We may be at a moment where it is genuinely difficult to know whether we are approaching a time where Artificial General Intelligence (AGI) becomes real, or whether we are once again witnessing a false dawn or new AI winter (Hendler 2008). Even amongst

some of the currently noted authorities there is little consensus. Roboticist Rodney Brooks for instance thinks that Artificial General Intelligence is a long way off, while Stuart Russel (2019) claims we may be only one major innovation away from human level performance. A new wave of AI ethicists (Coeckelbergh 2020; Taddeo and Floridi 2018) argue that many of the scenarios entertained about AGI are closet science fiction that distract us from clear thinking about the real ethical dilemmas posed by the more limited but real AI of today. Yet, with so much disagreement among experts, it remains generally unclear what the status of AI is today. At the time of going to press, the limits of what AI can do seem to be being challenged on an almost daily basic. It it difficult to put away the existential questions posed by AI.

If we have, up until now in this introduction, followed Freud in framing the challenges to the self-conception of humanity in terms of a peculiarly Western self-image, it is time to acknowledge that AI's entrance into our social life, culture and especially the mental schemes with which we conceive of mind is truly a global phenomenon. It is true that there are significant differences between the way that the major economically advanced nations have reflected on AI and how it affects public consciousness. One reason for this may be that different tacit assumptions are embedded in the folk philosophical systems and the intellectual heritage of conceptual schemes of different peoples, that form the background structure of much of thought (Baggini 2018). Different cultures think about the possibility of AI in different ways which are partly shaped by traditional ideas about the role and nature of human beings as well as different cultural experiences. As important as these cultural differences may be, the challenge that AI poses to the human self-image cuts across culture in many ways.

It is, for instance, frequently reported that Japan has much less cultural anxiety about the adoption of AI and the application of advanced robotics. Indeed, Japan has been considered a world superpower in robotics for something like the last 30 years (Smart et al. 2019). Traditional Japanese culture has arguably put much less stress on the idea of human beings dominating nature. Rather according to the still state religion of Shinto, human beings are traditionally seen as a part of nature (Ito 2018). We should note too a significantly different tradition of thinking about "cyberculture" through the prism of Manga and related science fiction that now deeply influences the West.

In 2017, Japanese Prime Minister Shinzo Abe observed that "Japan has no fear of AI. Machines will snatch away jobs? Such worries are not known to Japan. Japan aims to be the very first to prove that growth is possible through innovation, even when a population declines" (Kharpal 2017). Economic factors are thus also likely playing an important role. Rodney Brooks (2002) noted almost two decades ago that the declining population trends and strict immigration controls in Japan created a market for robot caregivers that could look after the sick and elderly. How these factors will influence the burgeoning international competition in taking maximum economic advantage of AI is difficult to interpret. Yet even in Eastern societies the challenge of AI to human self-image has been widely registered.

Perhaps the signal moment in registering the global impact of AI on the world's public consciousness was the 1–4 defeat by Korean 9 dan rank champion Lee Sedol

by Google Deep-Mind's program AlphaGo in a series of matches in March 2016. AlphaGo, the documentary movie, fantastically illustrates the shock and melancholy of encountering an AI that can play a game that is so associated with human intelligence and sense of self (Kohs 2017).

The successor program to AlphaGo, AlphaGo Zero already exhibits a highly restricted form of general intelligence in the sense that it can be applied to an open-ended set of games including Chess and Shogi (Silver et al. 2017). An important aspect of this fact is that the algorithms are achieving greater generality exactly as the AI approach becomes less specialized. AlphaGo Zero achieved its success in part by replacing many of the hand-coded heuristics with a more general Monte Carlo tree search algorithm. The algorithm can be viewed as a sort of fusion of early search space based AI, with sophisticated neural network techniques.[38] At the time of this book going to press the basic technology of Alpha Zero has been applied in a new form to the protein folding problem, which appears to have moved a long way toward a solution to a 50 year old scientific grand challenge (Callaway 2020).

Another contrast case that is worth briefly reflecting upon is that of China. The Chinese state has already unleashed massive spending in AI following its State Plan in AI Factors including the SmileToPay program, which uses face recognition algorithms to validate a citizen's identity through the *Social Credit Plan* (Smart et al. 2019). China has moreover already made plans for significant further investment. A major potential advantage for China's bid to become a world leader in AI technology is the Chinese Communist Party's ability to centralize and control their citizens' data. One report also claims that in contrast to the Western debate, "there is little to no discussion of issues of AI ethics and safety in China." (Ding 2018). (This can, of course, be contrasted with the "Coordinated Plan on AI" being published by the EU at the time of going to press that substantially focuses on the risks, legal framework and ethics of AI as developed and used within EU territories).

The Chinese model is just one model – albeit a very important one in today's technological landscape – and other models of how AI will become embedded in human societies are surely possible. Globally, the adoption and deployment of AI is developing rapidly along different paths in different legal and political frameworks. Yet even with Light AI, the space that is currently being adopted and its effect on human institutions is very various. We have discussed here how even weak AI, smart, rather than truly intelligent gadgetry can, if we come to deeply depend upon it, transform human cognition. If AI is to become a global competition it may be that human minds will be transformed in very different ways across the globe. This is the problem with dismissing certain research directions as science fiction. There is little doubt that we are already relying on a multitude of forms of AI that are changing the human cognitive profile. As we do, so we are entering a new realm of *Mind Design Space* in which conscious human intervention appears possible (Clowes 2020b), but also largely unconscious and haphazard experimentation is highly likely. Foreclosing

[38] For a nice description analysis of how these algorithms were developed see Somers (2019).

our imagination of this space now risks endangering our ability to shape it, now and in the future.

The Mind-Technology Problem then seems to be being posed in ever ramifying spheres wherever we consider the nature of the human mind and its differences. One way of resolving this problem, suggested by Mazlish, is to admit that the differences between ourselves and the intelligent machines we are creating as not one that can be ultimately maintained. Mazlish sees this – extending a certain Marxist tradition – as a process of overcoming our alienation from our artefacts. Others, see the ways in which we now engage with AI artefacts, as though they were persons, as itself a profound form of alienation (Turkle 2011).[39] For the time being, at least the sorts of AI technology with which we are interacting exists more towards the *light* end of the spectrum and it seems important that we learn to think about, develop policies around and live with this sort of AI. Yet, the technology is itself undergoing a rapid process of evolution and what appears to be science fiction today may not be tomorrow. AI that can, for instance, be tailored to write an article about AI for the Guardian newspaper changes the way human beings are likely to apply and think about their own intelligence. The matter is controversial (cf. Floridi and Chiriatti 2020b) but the future of algorithms like GPT-3 is rapidly shifting our ideas of what AI systems make possible.[40]

A third way between the various forms of alienation is to acknowledge the radical openness of the problem.[41] The scope and possibilities of AI are currently unknown, but AI is already radically changing the nature of human cognition. To shape this process of change we need to deeply engage with the fundamentally ethical and normative questions of what kinds of minds we both want to have, and want to create. What kind of beings we want to be. It is not clear yet whether we have too much, or too little "science fiction" to help us do this job. But, fully engaging with the range of ideas which will help us shape the future is a vital endeavour. In what follows we will trace how some of the contours of the Mind-Technology Problem are explored in the rest of this book and invite the reader to further explore these vital questions with us.

[39] For a critique and discussion of Turkle's *Alone Together* see Clowes (2011). For a detailed discussion of what it would take for an artificial being to count as a person see Clowes (2020a).

[40] In fact, GPT-3 is yet another application of the deep learning model. Its scope, at least at the time of going to press, has so far been defined by its major successes in unexpected areas. Quite what its limits are, is so far unknown.

[41] See Steve Fuller's discussion of these themes and how they intersect with the questions of posthumanism and transhumanism (Fuller 2011) and the concluding chapter from Georg Theiner for further reflection on this topic.

1.3.1 Computational Technology and Emergence in Mind

The first section of the book deals with the metaphysics of the mind, especially in relation to consciousness and agency, and whether they might be realized or indeed emerge in artificial cognitive systems. Here the questions focus on how we should conceive of the basic properties of complex human-like minds and what sort of processes can be used to account for them. This means, under what circumstances can mental properties – especially human-like mental properties – arise in artificial, natural or indeed hybrid systems. Should we really apply concepts such as human agency and consciousness to artificial minds and what prospects are there that such properties could survive intact among our artefactual creations? Reciprocally, how might our notions of mind, consciousness and agency change as we rethink them through the prism of our relationship with technology?

Mark Bickhard's paper, *Emergent Mental Phenomena*, begins this section with a discussion of whether or not mental phenomena – and especially consciousness – could emerge in artificial systems. Bickhard thinks that this should be possible in principle, but not with our current dominant computational technology. To set up his argument, in a first step, the author discusses the conditions of the possibility of emergence of mental phenomena in dynamic far-from-equilibrium systems. He argues that the standard particle or substance metaphysics does not allow for emergence at all. Therefore, Bickhard turns to quantum field theory which allows for a process ontology (Bickhard 2009; Ney and Albert 2013) and thus a possible theorization of emergence. To capture this dynamic nature of mental phenomena, the author introduces the notion of normative emergence. This type of emergence is grounded in normative functions which need to be established to maintain the organization of far from equilibrium systems, e.g., biological systems. In a second step, Bickhard builds upon the notion of *representing* found in Piaget (1954) as a model of consciousness that is non-unitary: composed of a primary non-conceptual form of awareness, and of a reflective aspect. Such consciousness Bickhard argues could, in principle, emerge from technology, but it could not emerge from our current digital computational technology.

Keith Frankish's wide-ranging paper, *Technology and the Human Minds,* is next, and continues this investigation of the relationship between reflective and what Frankish calls intuitive cognitive processes and re-interprets them in terms of the dual-process theory of the human mind and consciousness. Frankish builds his account upon the distinction between two types of processing systems found in the human brain (Evans 2010; Kahneman 2011).[42] On this view, human cognitive processing systems can be divided into the fast, unconscious, automatic, and evolutionarily old system 1 processes, and the slow, deliberative and potentially conscious system 2 processes. For Frankish, system 2 processes compose what he calls the

[42] See Frankish's paper in this volume for a more extensive bibliography on the two system view. For a deeper analysis of its background and how it has intersected with the history of philosophy see Frankish (2010).

virtual mind. These more recently created systems are largely products of human culture and often loop through the environment in the manner that Dennett (1991) describes as autostimulation.[43] This re-evaluation of system 2 processes, if correct, has major implications for cognitive enhancement and AI. Frankish argues that some forms of cognitive enhancement are relatively easily accomplished, as, in a sense, system 2 processes are all cognitive enhancements of more basic systems. The enhancement of system 1 processes, by contrast, would be much more difficult implying profound biological and or developmental intervention. Creating an AI, Frankish claims, is far more challenging than often assumed, especially if it follows the AGI model. This is because human conscious minds are not general intelligences either, but piecemeal culturally constructed systems build upon type 1 processes. Top-down approaches to AI especially would run into trouble here. Frankish's model implies the possibility of the open-ended development of cognitive enhancement, since the type 2 systems that are associated with consciousness, are not only virtual in Frankish's sense, but may be dependent upon, or indeed partly be embodied by, already existing technological systems.

Danielle Swanepoel's paper, *Does Artificial Intelligence have Agency?* turns our attention to a different but central property of minds – both natural and perhaps artificial – namely agency. As the title states, the objective of this chapter is to explore whether or not AI has agency. To do so, the author introduces some of the most celebrated accounts of agency, namely those developed by Harry Frankfurt (1971), Michael Bratman (2007), Christine Korsgaard (2009) and J. David Velleman (2009). Swanepoel uses these accounts to extract the essential features for agency and unite those in her own composite approach: *Common-Ground Agency (CGA)*. She settles upon four features, namely, *deliberative self-reflection, awareness of self in time, critical awareness of environment* and *norm violation*. Swanepoel then uses these conceptual categories to assess the possibility for agency in AI systems. In a first step, she admits that these features are approached from a possibly unwarranted anthropocentric perspective and makes them more AI-friendly. Still, or so the author argues, none of these features can currently be implemented in a way that would lead us to conclude that AI systems have agency because of their computational nature. Swanepoel notes that there is an interesting "correlation" between her four features of CGA and phenomenal consciousness. The paper discusses two options for this correlation: on the one hand, CGA's features do not require phenomenal consciousness, and therefore this is not the reason why AI fails to have agency. This is the case, since as far as we know AI currently does not instantiate consciousness. She discards this possibility on the ground that intuitively phenomenal consciousness plays an important role for establishing agency. On the other hand, if CGA's features require consciousness, then a theory that explains phenomenal consciousness, but cannot show that AI possesses it, might show why AI does not comply with the features of CGA.

[43] See Chapter 7 of Dennett's (1991) *Consciousness Explained* for a detailed discussion of autostimulation and why this cultural invention may hold the key to unlocking the latent powers of the human brain to new purposes.

In the final chapter of this section, *Consciousness: Philosophy's White Whale*, Gerald Vision further explores the metaphysics of consciousness in the context of emergence. He holds that at the intersection of mind and technology some new metaphysical implications may arise for phenomenal consciousness, particularly the question of whether phenomenal consciousness can arise in computational systems. To do so, Vision introduces Intel-Mary, a version of Mary the neuroscientist who was brought up in a black and white room from the thought experiment originally conceived of by Frank Jackson (1982). Vision's Intel-Mary however undergoes a piecemeal brain replacement until a large amount of her brain is artificial. Vision asks us to consider at which point in the replacement procedure might we doubt that Mary is still sentient? To set up the discussion about artificial Mary, Vision first analyses the plausibility of two metaphysical views, namely monism and panpsychism in relation to emergentism about phenomenal consciousness. Important for Vision is to first disarm the often-held charge that the emergence relation is brute and *ad hoc* by explaining that brute relations always occur in the sciences, even in physics. The implication for Vision is that in the case of monism/panpsychism the concept of Intel-Mary would not make sense since this view entails protoconsciousness as an essential basis. Emergentism, however, allows one to ask whether Intel-Mary is still phenomenally conscious and hence opens the door to ask the question of whether machines can be phenomenally conscious.

1.3.2 The Metaphysical and Technological Presuppositions of Mind Uploading

The second section of the book discusses the problem of mind-uploading and digital immortality, especially how this question interlocks with our understanding of the capabilities of current and near future computational technology. The contributions assess the problem of whether or not we can upload human minds to computers and whether or not the result of the uploading process does in some sense resemble or is the very same person as the putatively uploaded person. Implicitly this section also implies a reassessment of the way that computational technology can provide a proper conceptualization and indeed technical substrate for the minds of persons.

Gualtiero Piccinini's paper, *The Myth of Mind Uploading*, begins this section by challenging the idea that mind uploading is a serious possibility. Piccinini notes that many people think that mind uploading is feasible because, thanks to computational concepts in vogue, it is often held that the mind is like a software program running on our brain (computer functionalism) (Piccinini 2010), while, at the same time, underestimating the difficulties in simulating brains. Consequently, many assume that computer-based brain simulations – where the brain is simulated within a digital computer – or the continuous replacement of the biological brain by machine parts is possible in principle. Piccinini, however, thinks that this is unlikely. Firstly, constructing an accurate brain simulation or replacing a natural brain by neural

prosthetics is not enough to upload our mind. To represent the mind, the uploaded brain has to represent a particular, individual brain including emotions, idiosyncrasies, personality, evolution over time (Paul Smart, this volume, will pick up on this idea in Chap. 9). According to Piccinini, however, neuroscience is not in the business of investigating individual brains. It rather examines the general structure of all brains. Further, it is very unlikely that brain simulations or replacement scenarios would exhibit consciousness. According to the author, this is due to the fact that we do not know what the physical basis for consciousness is (Schneider and Corabi, this volume, make related points). Finally, there is still the issue of survival. In this context, Piccinini explicitly avoids becoming entangled with the personal identity debate. For him, the question of survival can be settled by noting that for one to survive the uploading process, it is necessary that there are no further copies of oneself. After discussing some examples of what a duplication process is, Piccinini concludes that a brain simulation is actually just a form of duplication. Only the brain replacement scenario may be a serious candidate to survive mind uploading. However, since we do not know what the physical basis of consciousness is and since neuroscience does not give us the specifics about a particular, individual brain, Piccinini discards the idea that the mind is simply a software running on our brain. As a consequence, for him mind uploading is a myth.

The paper from Corabi, *Cyborg Divas and Hybrid Minds*, extend previous work on mind-uploading (Corabi and Schneider 2012, 2014; Schneider 2009, 2019) that questions whether the vision of the mind presupposed by many advocates of the possibility of mind-uploading is conceptually coherent. In their work, they have questioned the interpretation of the software view of mind that seems to be held by many uploading advocates (Bostrom and Sandberg 2008), and especially, they have challenged the idea of survivability. That is, even if some apparent computational successor entity can be created through the uploading process, they give us reasons to challenge the idea that this entity would really be *you*.[44] The paper begins with a discussion of the extended mind thesis (EMT) (Clark and Chalmers 1998) and the authors's argue that certain contemporary neural prosthetics should also be considered as falling under EM's theoretical framework. Since, neural prosthetics are already being developed, this can, according to Schneider and Corabi, make the EM thesis and even the extended consciousness (EC) thesis a testable hypotheses. Since the cognitive basis of consciousness is decisive for mind-uploading, it is important to know whether brain chips can actually form part of this basis or not. Schneider and Corabi, then turn to the case of mind-uploading, which they consider a form of radical enhancement. In their paper, the authors revisit a classical scenario of mind-uploading, namely instantaneous destructive mind-uploading and remind us why they think a person cannot survive this process (Corabi and Schneider 2014; Piccinini this volume). In what follows, Schneider and Corabi analyse the key potential counterexample to their argument about the impossibility of

[44] Many of these themes have been developed in deep and exciting new ways in Schneider's book *Artificial You: AI and the Future of Your Mind* (Schneider 2019).

mind-uploading, namely the challenge from EM (e.g., Clowes and Gärtner, this volume). Here, the upload process is essentially different to what has been considered so far. Most importantly, since the EM thesis and especially the EC thesis hold that the basis of the mind and consciousness may be extended – and since we are already partially extended by our current digital technology – it may be concluded that we are already partially uploaded. This challenge puts in doubt the idea that we cannot survive the mind-uploading process. In the final sections of the paper, Schneider and Corabi, deal with this challenge and argue that it is misguided at best, especially since the EC thesis is itself highly doubtful.

In the next chapter, Clowes and Gärtner in their paper, *Slow Continuous Uploading*, directly take up the challenge of Corabi and Schneider's (2014) case against the survivability of mind-uploading. They begin with the metaphysical question over whether a person should be construed as a substance or object-like entity and how this frames the argument that the uploading process cannot be the continuation of the original person. From this perspective, objects – and therefore persons – cannot entertain the temporal and spatial discontinuities which the uploading process assumes. Clowes and Gärtner, however, point out that this is only the case in standard or "vanilla uploading scenarios" such as instantaneous destructive uploading or brain replacement, which are also negatively evaluated by Piccinini (this volume). As a consequence, they examine a different route to uploading which gets much of its theoretical heft from the notion of the extended mind (Clark and Chalmers 1998). The idea here is that a form of partial uploading may already be happening through our habit of using social media systems and lifelogging technologies to upload ever more digital traces of ourselves into computational media, and then crucially, by maintaining deep and ongoing interactions with these systems. Examining this *slow continuous uploading* (SCU) scenario, Clowes and Gärtner ask, which aspects of personhood could persist through such a proces? The paper, then, directly takes up the challenge from Schneider and Corabi's paper (this volume) that argues that a partial upload could confer neither survivability or continuation, since it would lack core cognitive and especially core conscious features of an individual's mind. Clowes and Gärtner address this worry by questioning a key assumption of this argument, which holds that there is a clear division between core and peripheral cognitive and conscious aspects of the mind. They remind us that in Dennettian (Dennett 1991) or Clarkian (Clark 2006) views of mind, a clear core to consciousness and the self is difficult to establish. Finally, they admit that even SCU, when understood as maintaining some psychological continuities, has to deal with serious challenges about other forms of discontinuity. However, Clowes and Gärtner argue, that these discontinuities may not be more profound than the transitions that allow a form of continuance in nature, such as the life cycle of a butterfly that, despite very different incarnations throughout its life-stages, still counts as one and the same individual.

Paul Smart's paper, *Predicting Me: The Route to Digital Immortality?*, continues the theme of SCU, but with a focus on the mechanics of how current technologies and theories of the functioning of the brain – especially predictive processing (Clark 2015b; Friston 2008; Hohwy 2013) – might be realized with in current approaches

to machine learning. Smart begins the paper with an analysis of deep learning techniques (e.g., Bengio 2009) and their surprising potential to be able to reconstruct deep patterns of causal influence in a variety of datasets. These machine learning techniques mirror some of the key assumptions of the predictive processing approach, namely the fact that the architecture of the brain should be considered as hierarchical and multi-layered, and that the brain is characterized by the ability to build and use generative models to capture the structure of the world. Assuming that the brain really is a prediction machine (Clark 2015b; Hohwy 2013), this implies that deep learning systems may be able to emulate the functions of the brain relevant to, e.g., personal identity by acquiring and implementing the right kind of generative models (Clark 2012). Smart's approach to SCU seems to circumvent some of Piccinini's objections (this volume) by considering a new model of how the functional properties of a brain might be digitally encoded in a non-destructive manner. By locating the question of digital immortality against the background of the predictive processing approach to the mind, and especially deep learning technology, Smart offers us a series of detailed scenarios whereby long-term interaction with predictive technologies might accomplish a form of uploading. We leave it to the reader to decide how successful Smart's suggestions are for circumventing the more general line of the argument found in Piccinini's paper about the technological feasibility of a form of mind-uploading and also whether such techniques of SCU could ever constitute a form of personal survival or continuation (as discussed in the papers by Schneider and Corabi and Clowes and Gärtner in this volume).

One central background context of this section is Susan Schneider's recent book *Artificial You: AI and the Future of Your Mind* (Schneider 2019). As this volume shows, some of the most important questions for the future are likely to arise where the idea of the extended mind intersects with concepts of what it is to be a person or a self. Some of the implications and indeed the deep history of these questions are examined in the third section of the book.

1.3.3 The Epistemology, Ethics and Deep History of Mind Extension

The third section of the book discusses the extended mind thesis in the context of twenty-first century technology and the deep history of Western concepts of the mind and uses this perspective to attempt to rethink central mentalistic concepts such as agency, knowledge and introspection. Contributions in this section deepen our consideration of the extended mind, but expands the horizon of the implications into a focus on the ethical and epistemic implications of mind extension and more generally on what the implications are of being human in a time where our minds are to an ever greater extent extended by a growing range of "smart" tools and systems.

One domain where the ethical issues become dramatized is the use of drone technology by the military. Marek Vanžura, in *What it is like to be a drone operator? Or, remotely extended minds in war*, examines the case of why drone operators working in the military context may suffer from post-traumatic stress disorder (PTSD). One major reason for the deployment of drones is that they enable a soldier's apparent risk-free participation in missions in war zones. According to Vanžura, however, this may only apply to the physical risk for the pilots of drones. The reason that it is physically risk-free is that such operations are conducted from afar, sometimes by pilots outside the country of engagement in order to mitigate pilot risk. But, according to psychological studies (Chappelle et al. 2014a, b), even though drone operators do not have to physically enter a zone of conflict, they are still prone to suffer from PTSD. Vanžura explores this circumstance with the aid of the extended mind thesis. For him, the fact that drone operators suffer from PTSD can be explained in terms of how cognitive processes, and indeed the minds of the operators, are extended to the drone. Through the drone interface, there is a two-way reciprocal interaction between the operator and the drone, i.e., operators perceive through the drone's sensory apparatus (e.g., infra-red cameras), and drones follow the change direction on the operator's command. The operator engages in real world manipulations based on the drone technology. According to Vanžura's hypothesis, the operator's cognitive processes are extended into the geographic areas in which the drone's weapons have effects, exposing drone operators to the psychological effects of war.

In the second chapter of this section *Extending Introspection*, Lukas Schwengerer argues for the possibility of extended introspection in circumstances of import to our contemporary technological setting. This idea is based on the extended mind thesis. He claims that extended introspection should not be thought of as a variation of traditional theories of introspection (Schwitzgebel 2019), but still agrees to the core claim that the obtained knowledge is privileged. To set up his discussion, he uses the classical extended mind scenario of Otto, which he later expands to the scenario of Otto++ (Smart 2018). Otto++ does not note things in his notebook, but has the requisite information stored on his personal server which he can access by internet technology such as a smart phone, smart watch, augmented glasses, etc. Assuming the extended mind thesis from the original Otto thought experiment is tenable, Schwengerer thinks that there are two conflicting intuitions about extended introspection. On the one hand, Otto's self-ascriptions seem to be based on directly detecting his own beliefs – something that also happens in the case of traditional introspection. On the other hand, Otto clearly employs evidence that is also accessible to someone else. After discarding the idea that extended introspection is just another form of introspection or mind reading, Schwengerer argues that extended introspection is based on a particular set of epistemic rules that only apply to these extended cases; it is this that guarantees privileged access. Finally, the author argues that his account is not only valid in the simple and limited scenario of Otto, but also in the case of Otto++. Schwengerer concludes that using twenty-first century cloud technology satisfies all constraints necessary for his account of extended introspection to be further generalized.

In a third contribution to this section, Gloria Andrada's paper, *Epistemic Complementarity: Steps Towards a Second Wave Extended Epistemology*, discusses a new way to tackle extended epistemology (Carter et al. 2018). Her paper relates directly to the new conditions we find for discussing the nature of knowledge in the age of cloud technology. Andrada argues that up until now extended epistemology views are based on the first-wave approach to the extended mind (Clark and Chalmers 1998). In her opinion, however, this framework leads to inadequate interpretations of the needed epistemologically valuable extended cognitive processes. Therefore, she proposes an alternative view which is modelled upon a second wave discussion of extended cognition (Menary 2010; Sutton 2010). The second-wave approach to the extended mind aims beyond the parity principle and reliance on coarse-grained functional similarities between intracranial and extended cognition. According to the complementarity principle (Sutton 2010) we tend to incorporate artefacts into our cognitive routines when those artefacts afford some cognitive advantage over already existent intra-cranial mechanisms.[45] Second-wave approaches to the Extended Mind, among other advantages, allows for the possibility of conceiving of types of cognition that may arise in the context of novel technologies and novel uses of technology.[46] For extended epistemology then, the key element is not epistemic parity, i.e., if an epistemic condition is valid for intracranial cognitive processes, it should also count for extended cognitive processes. Rather Andrada argues, we should be guided by an epistemic complementarity principle, i.e., epistemic validity depends on the interaction of the embodied knower, the properties of the technological artifact and the socio-cultural environment. Taking this seriously involves the complex challenge of analysing how our new technologies may be transforming the nature of individual and group cognitive epistemic abilities.

In this volume's final paper, Georg Theiner's contribution, *The Extended Mind: A Chapter in the History of Transhumanism*, returns us to the questions with which we began this discussion, namely the deep conceptual history of how mind and technology relate and how this might shape our future. The chapter develops a unique and challenging position on the mind-technology problem by situating the extended mind thesis and its relationship to transhumanism against the background of the Christian tradition. The paper sets the work of Andy Clark (2003, 2008), and especially the theory of the extended mind, in the deep historical context of the Christian view of human nature and embodiment. Theiner analyses, in detail, Steve Fuller's (2011) archaeology of the concepts of transhumanism and posthumanism within an essentially Christian interpretation of the nature of mind. Theiner's thought-provoking claim is that the extended mind thesis can be understood as a continuation of the Christian doctrine that human beings are built in the image of God dressed up in earthly clothing. According to the author, Clark's vision of humanity as natural born cyborgs is itself a form of transhumanism and a materialist account of how humanity can transcend itself to become in several senses

[45] See also Clark's discussion of the principle of ecological assembly (see Clark 2008).

[46] See further discussion in Clowes (2015).

God-like. Theiner's historical account sheds new light on the posthuman/transhuman debate and offers an original take on the special nature of the human mind, and how current exploration of the boundaries between mind and technology continue deep tendencies in the Western understanding of human nature in unexpected ways.

1.4 The Mind Technology Problem and the Future of Philosophy

The mind-technology problem, as we have presented it here, has two distinct stages. The first stage has its roots in the theoretical account of computation first developed by Alan Turing in the darkest days of the second world war. It was the computational model of mind and the attendant informational revolution that led to a new way of conceiving of minds and, in consequence, the place of human beings in nature. Its distinctive philosophical contribution was functionalism and the computational model of mind which ultimately provided a way of superseding the mind-body problem as it had developed from Descartes times. We have focused here on the mind-technology problem as a distinctly different constellation of problems about mind from those formulated in the early modern period. We present it to the reader as a new sort of Gestalt: a new way of arranging familiar problems in order to pursue novel solutions. With the mind-technology problem the focus shifts from how minds, conceived as ethereal substances, can *interact* with matter, to what, if anything, makes minds, cognitive processes, and indeed human intelligence special in a physical universe at all. The distinctive problems in the era of the mind-technology problem include: What properties of mind may be enabled, transformed or extended by technology? What properties of mind may be diminished, outsourced or curtailed? Is human agency being primarily constrained or enabled by our encounter with 21st Century technology and especially by our interaction with AI? How might the nature of human agency, memory, knowledge, responsibility and consciousness be changed through this interaction? These can all be viewed as problems of where our minds stop, and our artefacts begin. Deciding the limits of mind seem to recast the nature of the other philosophical problems around it.

The second phase of the mind-technology problem has been underway now for at least 10 years as the human race now spends a sizable proportion of its time interacting with, speaking to, and having our everyday lives structured by artificially intelligent systems. It is presents itself as not so much a series of metaphysical philosophical questions, but as a series of ethical and practical challenges. We want to re-expose the philosophical questions at the roots of this practical encounter. As we come to cohabit with AIs, the nature of human cognition and our conception of own minds is undergoing a radical transformation. Sometimes this change is in the background and scarcely noticed. This cannot be good. This is not science fiction but is the nature of the times we are living through. A central philosophical task of the coming decades will be to make sense of and shape this radical conceptual change

and with it attempt to promote the sorts of futures we want. We welcome you, dear reader, to the further pressing consideration of the mind-technology problem.

Acknowledgements Robert W. Clowes's work is supported by FCT, 'Fundação para a Ciência e a Tecnologia, I.P.' by the Stimulus of Scientific Employment grant (DL 57/2016/CP1453/CT0021) and personal grant (SFRH/BPD/70440/2010).

Klaus Gärtner's work is endorsed by the financial support of FCT, 'Fundação para a Ciência e a Tecnologia, I.P.' under the Stimulus of Scientific Employment (DL 57/2016/CP1479/CT0081) and by the Centro de Filosofia das Ciências da Universidade de Lisboa (UIDB/00678/2020).

This work is endorsed by the FCT project "Emergence in the Natural Sciences: Towards a New Paradigm" (PTDC/FER-HFC/30665/2017).

References

Armstrong, D. M. (1980). *The causal theory of the mind*.

Armstrong, D. M. (1983). *The nature of mind and other essays*.

Baggini, J. (2018). *How the world thinks: A global history of philosophy*. Granta Books.

Barkow, J. H., Cosmides, L., & Tooby, J. (1992). *The adapted mind: Evolutionary psychology and the generation of culture*. New York: Oxford University Press.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and trends® in Machine Learning, 2*(1), 1–127.

Benzon, W. L. (2020). GPT-3: Waterloo or Rubicon? Here be dragons. *Here be Dragons (August 5, 2020)*.

Bickhard, M. H. (2009). The interactivist model. *Synthese, 166*(3), 547–591.

Boden, M. A. (1977). *Artificial intelligence and natural man*. Hassocks: Harvester Press.

Boden, M. A. (1990). *The creative mind: Myths and mechanisms*. London: Sphere Books Ltd.

Boden, M. A. (2006). *Mind as machine: A history of cognitive science two-volume set*. Oxford: Oxford University Press.

Bostrom, N. (2014). *Superintelligence*. Dunod.

Bostrom, N., & Sandberg, A. (2008). *Whole brain emulation: A roadmap*. Lancaster University. Accessed 21 Jan, 21, 2015.

Bratman, M. (2007). *Structures of Agency: Essays*. Oxford: Oxford University Press.

Brooks, R. (1990). Elephants don't play chess. *Robotics and Autonomous Systems, 6*, 3–15h.

Brooks, R. (1991a). *Intelligence without reason.* Paper presented at the International Joint Conference on Artificial Intelligence.

Brooks, R. (1991b). Intelligence without representation. *Artificial Intelligence, 47*, 139–160.

Brooks, R. (2002). *Robot: The future of flesh and machines*. Cambridge, MA: Allen Lane: The Penguin Press.

Brooks, R., Breazeal, C., Marjanovic, M., Scassellati, B., & Williamson, M. W. (1999). *The cog project: Building a humanoid robot*.

Bruner, J. S. (1956). Freud and the image of man. *American Psychologist, 11*(9), 463.

Callaway, E. (2020). *'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures*.

Carr, N. (2008). Is Google making us stupid? *Yearbook of the National Society for the Study of Education, 107*(2), 89–94.

Carr, N. (2010). *The shallows: How the internet is changing the way we think, read and remember*. London: Atlantic Books.

Carter, J. A., Clark, A., Kallestrup, J., Palermos, S. O., & Pritchard, D. (2018). *Extended epistemology*. Oxford: Oxford University Press.

Castells, M. (1996). *The information age: Economy, society and culture* (3 volumes). Oxford: Blackwell, 1997, 1998.

Castro-Caldas, A., Petersson, K. M., Reis, A., Stone-Elander, S., & Ingvar, M. (1998). The illiterate brain. Learning to read and write during childhood influences the functional organization of the adult brain. *Brain, 121*(6), 1053–1063.

Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies, 2*(3), 200–219.

Chalmers, D. (2002). Consciousness and its place in nature. In S. Stich & T. Warfield (Eds.), *Blackwell guide to the philosophy of mind*. (Reprinted from: online at Chalmers website. http://consc.net/papers/nature.html).

Chalmers, D. (2015). Panpsychism and panprotopsychism. In T. Alter & Y. Nagasawa (Eds.), *Consciousness in the physical world: Perspectives on Russellian Monism*. Oxford: Oxford University Press.

Chappelle, W. L., Goodman, T., Reardon, L., & Thompson, W. (2014a). An analysis of post-traumatic stress symptoms in United States Air Force drone operators. *Journal of Anxiety Disorders, 28*(5), 480–487.

Chappelle, W. L., McDonald, K. D., Prince, L., Goodman, T., Ray-Sannerud, B. N., & Thompson, W. (2014b). Symptoms of psychological distress and post-traumatic stress disorder in United States Air Force "drone" operators. *Military Medicine, 179*(Suppl_8), 63–70.

Clark, A. (2003). *Natural born cyborgs: Minds, technologies and the future of human intelligence*. New York: Oxford University Press.

Clark, A. (2006). Soft selves and ecological control. In D. Spurrett, D. Ross, H. Kincaid, & L. Stephens (Eds.), *Distributed cognition and the will*. Cambridge, MA: MIT Press.

Clark, A. (2008). *Supersizing the mind*. New York: Oxford University Press.

Clark, A. (2012). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–204.

Clark, A. (2015a). *Predicting peace: The end of the Representation Wars Open MIND: Open MIND*. MIND Group: Frankfurt am Main.

Clark, A. (2015b). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford: Oxford University Press.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*, 10–23.

Clowes, R. W. (2011, Monday 31st October). Electric selves? Review of alone together: Why we expect more from technology and less from each other, by Sherry Turkle. Culture Wars.

Clowes, R. W. (2013). The cognitive integration of E-memory. *Review of Philosophy and Psychology*, (4), 107–133.

Clowes, R. W. (2015). Thinking in the cloud: The cognitive incorporation of cloud-based technology. *Philosophy and Technology, 28*(2), 261–296.

Clowes, R. W. (2017). Extended memory. In S. Bernecker & K. Michaelian (Eds.), *Routledge handbook on the philosophy of memory* (pp. 243–255). Abingdon/Oxford: Routledge.

Clowes, R. W. (2019). Immaterial engagement: Human agency and the cognitive ecology of the internet. *Phenomenology and the Cognitive Sciences, 18*(1), 259–279. https://doi.org/10.1007/s11097-018-9560-4.

Clowes, R. W. (2020a). Breaking the code: Strong agency and becoming a person. In T. Shanahan & P. R. Smart (Eds.), *Blade runner 2049: A philosophical exploration* (pp. 108–126). Abingdon/Oxon: Routledge.

Clowes, R. W. (2020b). The internet extended person: Exoself or Doppleganger? *Limité. Límite. Revista Interdisciplinaria de Filosofía y Psicología, 15(22).*

Clowes, R. W., Torrance, S., & Chrisley, R. (2007). Machine Consciousness: Embodiment and Imagination (editorial introduction). *Journal of Consciousness Studies, 14*(7), 7–14.

Coeckelbergh, M. (2020). *AI ethics*. MIT Press.

Copenhaver, R., & Shields, C. (2019a). General introduction to history of the philosophy of mind, six volumes. In R. Copenhaver & C. Shields (Eds.), *History of the philosophy of mind, Six Volumes*. Routledge.

Copenhaver, R., & Shields, C. (2019b). History of the philosophy of mind, Six Volumes.
Corabi, J., & Schneider, S. (2012). Metaphysics of uploading. *Journal of Consciousness Studies, 19*(7–8), 26–44.
Corabi, J., & Schneider, S. (2014). *If you upload, will you survive?* Intelligence Unbound*:* Future of Uploaded and Machine Minds, The, 131–145.
Dehaene, S. (2009). *Reading in the brain: The science and evolution of a human invention*. Viking Pr.
Dennett, D. C. (1978). *Artificial intelligence as philosophy and psychology Brainstorms*. Montgometry: Bradford Brooks.
Dennett, D. C. (1984). Cognitive wheels: The frame problem of AI. *Minds, Machines and Evolution*, 129–151.
Dennett, D. C. (1991). *Consciousness explained*. Harmondsworth: Penguin Books.
Dennett, D. C. (1996a). Facing backwards on the problems of consciousness. *Journal of Consciousness Studies, 3*(1), 4–6.
Dennett, D. C. (1996b). *Kinds of minds: Towards an understanding of consciousness*. New York: Phoenix Books.
Ding, J. (2018). Deciphering China's AI dream. *Future of Humanity Institute Technical Report*.
Dreyfus, H. L. (1972). *What computers can't do: A critique of artificial reason*. New York: Harper.
Evans, J. S. B. (2010). *Thinking twice: Two minds in one brain*. New York: Oxford University Press.
Fauconnier, G., & Turner, M. (2002). *The way we think: Conceptual blending and the mind's hidden complexities*. New York: Basic Books.
Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality*. Oxford: Oxford University Press.
Floridi, L. (2015). *The Onlife Manifesto: Being Human in a Hyperconnected Era*. Cham/Heidelberg/New York/Dordrecht/London: Springer.
Floridi, L. (2020). AI and its new winter: From myths to realities. *Philosophy & Technology*, 1–3.
Floridi, L., & Chiriatti, M. (2020a). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines, 30*(4), 681–694.
Floridi, L., & Chiriatti, M. (2020b). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 1–14.
Fodor, J. (1975a). *The language of thought*. Cambridge, MA: MIT Press.
Fodor, J. A. (1975b). *The language of thought*. New York: Harvard University Press.
Fodor, J. (2009). Where is my mind. *London Review of Books, 31*(3), 13–15.
Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy, 68*(1), 5–20.
Frankish, K. (2010). Dual-process and dual-system theories of reasoning. *Philosophy Compass, 5*(10), 914–926.
Freud, S. (1920). *A general introduction to psychoanalysis*. Createspace Independent Publishing Platform.
Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology, 4*(11), e1000211.
Fuller, S. (2011). *Humanity 2.0: Foundations for 21st century social thought*. London: Palgrave Macmillan.
Gallagher, S. (2001). The practice of mind. *Journal of Consciousness Studies, 8*(5–7), 83–108.
Gardner, H. (1985). *The mind's new science*. New York: Basic Books.
Gerken, M. (2014). Outsourced cognition. *Philosophical Issues, 24*(1), 127–158.
Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.
Goertzel, B., & Pennachin, C. (2007). *Artificial general intelligence* (Vol. 2). Springer.
Greenfield, S. (2015). *Mind change: How digital technologies are leaving their mark on our brains*. Random House.
Gregory, R. L. (1981). *Mind in science: A history of explanations in psychology*. Cambridge: Cambridge University Press.

Heersmink, R. (2016). Distributed selves: Personal identity and extended memory systems. *Synthese*, 1–17.
Heersmink, R. (2020). Varieties of the extended self. *Consciousness and Cognition, 85*, 103001.
Hendler, J. (2008). Avoiding another AI winter. *IEEE Intelligent Systems*, (2), 2–4.
Hohwy, J. (2013). *The predictive mind*. Oxford University Press.
Holland, O. (2003). Editorial Introduction. *Journal of Consciousness Studies, 10*(4), 1–6.
Hutto, D. D. (2008). *Folk psychological narratives: The sociocultural basis of understanding reasons*. The MIT Press.
Ihde, D. (1990). *Technology and the lifeworld: From garden to Earth*. Indiana University Press.
Ihde, D., & Malafouris, L. (2019). Homo faber revisited: Postphenomenology and material engagement theory. *Philosophy & Technology, 32*(2), 195–214. https://doi.org/10.1007/s13347-018-0321-7.
Ito, J. (2018). *Why westerners fear robots and the Japanese do not*. Wired.
Jackson, F. (1982). Epiphenomenal Qualia. *The Philosophical Quarterly, 32*, 127–136.
Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.
Kharpal, A. (2017). Japan has no fear of AI — It could boost growth despite population decline, Abe says. *cnbc.com*. Retrieved from https://www.cnbc.com/2017/03/19/japan-has-no-fear-of-ai%2D%2Dit-could-boost-growth-despite-population-decline-abe-says.html
Kim, J. (2006). *Philosophy of mind* (2nd ed.). Cambridge, MA: Westview.
Kind, A. (2018). The mind-body problem in 20th-century philosophy. Philosophy of mind in the twentieth and twenty-first centuries. *The History of the Philosophy of Mind, 6*, 1.
Knappett, C., & Malafouris, L. (2008). Material and nonhuman agency: An introduction. *Material Agency: Towards a Non-Anthropocentric Approach*, ix–xix.
Kohs, G. (Writer). (2017). AlphaGo. In G. Krieg, J. Rosen, & K. Proudfoot (Producer): RO*CO FILMS.
Korsgaard, C. M. (2009). *Self-constitution: Agency, identity, and integrity*. Oxford: Oxford University Press.
Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh*. New York: Basic Books.
Lakoff, G., & Johnson, M. (2003 [1980]). *Metaphors we live by*. Chicago: University of Chicago Press.
Laland, K. N., Odling-Smee, J., & Feldman, M. W. (2000). Niche construction, biological evolution, and cultural change. *Behavioral and Brain Sciences, 23*, 131–175.
Lanier, J. (2010). *You are not a gadget: A manifesto*. London: Allen Lane.
Loh, K. K., & Kanai, R. (2015). How has the internet reshaped human cognition? *The Neuroscientist*, 1073858415595005.
Luria, A. R. (1976). *Cognitive development: Its cultural and social foundations*.
Malafouris, L. (2010a). Grasping the concept of number: How did the sapient mind move beyond approximation. In I. Morley & C. Renfrew (Eds.), *The archaeology of measurement: Comprehending heaven, earth and time in ancient societies* (pp. 35–42). Cambridge: Cambridge University Press.
Malafouris, L. (2010b). Metaplasticity and the human becoming: Principles of neuroarchaeology. *Journal of Anthropological Sciences, 88*(4), 49–72.
Malafouris, L. (2013). *How things shape the mind: A theory of material engagement*. Cambridge, MA: MIT Press.
Malafouris, L. (2016). On human becoming and incompleteness: A material engagement approach to the study of embodiment in evolution and culture. *Embodiment in evolution and culture*, 289–305.
Mazlish, B. (1993). *The fourth discontinuity: The co-evolution of humans and machines*. Yale University Press.
McGeer, V. (2001). Psycho-practice, psycho-theory and the contrastive case of autism. How practices of mind become second-nature. *Journal of Consciousness Studies, 5–7*, 109–132.
Menary, R. (2010). Cognitive integration and the extended mind. In R. Menary (Ed.), *The extended mind* (pp. 227–244). London: Bradford Book, MIT Press.

Menary, R. (2014). Neural plasticity, neuronal recycling and niche construction. *Mind & Language, 29*(3), 286–303.
Milkowski, M. (2013). *Explaining the computational mind*. MIT Press.
Minsky, M., & Papert, S. (1969). *Perceptrons: An essay in computational geometry*. MIT Press.
Mithen, S. (1996). *The prehistory of the mind*. London: Thames Hudson.
Moor, J. (2006). The Dartmouth College artificial intelligence conference: The next fifty years. *AI Magazine, 27*(4), 87–87.
Moravec, H. (1988). *Mind children: The future of robot and human intelligence*. Cambridge, MA: Harvard University Press.
Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs: Prentice-Hall.
Ney, A., & Albert, D. Z. (2013). *The wave function: Essays on the metaphysics of quantum mechanics*. Oxford University Press.
Norman, D. A. (1993). *Things that make us smart* (*Defending human attributes in the age of the machine*). Addison-Wesley.
Ong, W. J. (1982). *Orality and literacy: The technologizing of the word*. London: Methuen.
Piaget, J. (1954). *The construction of reality in the child*. New York: Basic.
Piccinini, G. (2010). The mind as neural software? Understanding functionalism, computationalism, and computational functionalism. *Philosophy and Phenomenological Research, 81*(2), 269–311.
Piccinini, G. (this volume). The myth of mind uploading.
Postman, N. (1993). *Technopoly: The surrender of culture to technology*. New York: Vintage.
Putnam, H. (1967). Psychological predicates. *Art, mind, and religion, 1*, 37–48.
Putnam, H. (1980). The nature of mental states. *Readings in Philosophy of Psychology, 1*, 223–231.
Rumelhart, D. E., & McClelland, J. L. (1986a). *Parallel distributed processing: Exploring the microstructure of cognition* (Vol. 1). Cambridge MA: MIT Press.
Rumelhart, D. E., & McClelland, J. L. (1986b). *Parallel distributed processing: Exploring the microstructure of cognition* (Vol. 2). Cambridge, MA: MIT Press.
Russell, B. (1927). *The analysis of matter*. London: Kegan Paul, Trench, Trubner & Co.
Russell, S. J. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin Audio.
Ryle, G. (1949). *The concept of mind*. London: Hutchinson.
Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 3*(3), 210–229.
Sapolsky, R. M. (1997). *Junk food monkeys and other essays on the biology of the human predicament*. London: Headline.
Schneider, S. (2009). Mindscan: Transcending and enhancing the human brain. In S. Schneider (Ed.), *Science fiction and philosophy: From time travel to superintelligence* (pp. 260–276). Hoboken: Wiley- Blackwell.
Schneider, S. (2011). *The language of thought: A New philosophical direction*. MIT Press.
Schneider, S. (2019). *Artificial you: AI and the future of your mind*. Princeton: Princeton University Press.
Schwitzgebel, E. (2019). Introspection. In E. N. Zalta (Ed.), (Winter 2019 Edition ed., Vol. The Stanford Encyclopedia of Philosophy).
Searle, J. R. (1980). Mind, brains and programs. *Behavioral and Brain Sciences, 3*(3), 417–457.
Shelley, M. W. (2018). *Frankenstein: The 1818 text*. Penguin.
Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., et al. (2017). Mastering the game of go without human knowledge. *Nature, 550*(7676), 354–359.
Smart, P. (2018). Emerging digital technologies: Implications for extended conceptions of cognition and knowledge. In J. A. Carter, A. Clark, J. Kallestrup, S. O. Palermos, & D. Pritchard (Eds.), *Extended epistemology* (pp. 266–304). Oxford: Oxford University Press.
Smart, P. R., Heersmink, R., & Clowes, R. W. (2017). The cognitive ecology of the internet. In S. J. Cowley & F. Vallée-Tourangeau (Eds.), *Cognition beyond the brain* (2nd ed., pp. 251–282). Springer.

Smart, P. R., Madaan, A., & Hall, W. (2018). Where the smart things are: Social machines and the internet of things. *Phenomenology and the Cognitive Sciences*, 1–25.
Smart, P., Chu, M.-C. M., O'Hara, K., Carr, L., & Hall, W. (2019). Geopolitical drivers of personal data: The four horsemen of the datapocalypse.
Somers, J. (2019). How the artificial-intelligence program AlphaZero mastered its games.
Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science, 333*(6043), 776–778.
Sterelny, K. (2011). From hominins to humans: How sapiens became behaviourally modern. *Philosophical Transactions of the Royal Society B: Biological Sciences, 366*(1566), 809–822.
Sutton, J. (2010). Exograms and interdisciplinarity: History, the extended mind, and the civilizing process. In R. Menary (Ed.), *The extended mind* (pp. 189–225). London: Bradford Book, MIT Press.
Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science, 361*(6404), 751–752.
Tallis, R. (2004). *Why the mind is not a computer: A pocket lexicon of neuromythology* (Vol. 13). Imprint Academic.
Toffler, A. (1980). *The third wave* (Vol. 484). Bantam Books, New York.
Turing, A. M. (1937). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society, 2*(1), 230–265.
Turing, A. (1950a). Computing machinery and intelligence. *Mind, 49*, 433–460.
Turing, A. M. (1950b). Computing machinery and intelligence. *Mind, 59*(236), 433–460.
Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. New York: Basic Books.
Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind*. Cambridge, MA: MIT Press.
Velleman, J. D. (2009). *The possibility of practical reason*. Michigan Publishing, University of Michigan Library.
Vision, G. (2018). The provenance of consciousness. In E. Vitaliadis & C. Mekos (Eds.), *Brute facts* (pp. 155–176). Oxford: Oxford University Press.
Vygotsky, L. S. (1962). *Thought and language* (E. Hanfmann & G. Vakar, Trans.). Cambridge, MA: MIT Press.
Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge MA: Harvard University Press.
Vygotsky, L. S., & Luria, A. R. (1994). Tool and symbol in child development. In R. Van Der Veer & J. Valsiner (Eds.), *The Vygotsky reader*. Cambridge MA: Basil Blackwell.
Wegner, D. M., & Ward, A. F. (2013, December 1). The internet has become the external hard drive for our memories. Scientific American.
Weizenbaum, J. (1976). Computer power and human reason: From judgment to calculation.
Wilkes, K. V. (1984). Pragmatics in science and theory in common sense. *Inquiry, 27*, 339–361.
Wootton, D. (2015). *The invention of science: a new history of the scientific revolution*. Penguin, UK.
Zawidzki, T. W. (2013). *Mindshaping: A new framework for understanding human social cognition*. MIT Press.

Robert W. Clowes is senior researcher and coordinator of the Lisbon Mind and Reasoning Group at the Universidade Nova de Lisboa, Portugal. His research interests span a range of topics in philosophy and cognitive science, including the philosophy of technology, memory, agency, skills, and the implications of embodiment and cognitive extension for our understanding of the mind and conscious experience. He is particularly interested in the philosophical and cognitive scientific significance of new technologies, especially those involving the Internet and artificial intelligence and how these interact with human agency. His work has appeared in a variety of journals, including *TOPOI*, *Review of Philosophy and Psychology*, *AI & Society*, *Phenomenology and the Cognitive Sciences*, *Philosophy and Technology*, and the *Journal of Consciousness Studies*. He received his PhD from the University of Sussex.

Klaus Gärtner studied philosophy at the University of Regensburg. He obtained his PhD at the Instituto da Filosofia da NOVA (Universidade Nova de Lisboa). Currently, he is a researcher at the Departamento de História e Filosofia das Ciências and member of the Centro de Filosofia das Ciências da Universidade de Lisboa in the Faculdade de Ciências da Universidade de Lisboa. He is also a founding member of the Lisbon Mind and Reasoning Group. His research interests include philosophy of mind and cognitive science, philosophy of science, epistemology, and metaphysics.

Inês Hipólito is a postdoctoral fellow and a lecturer at the Berlin School of Mind and Brain (Humboldt Universität zu Berlin), and an affiliated member to the Neurobiology group at the Wellcome Centre for Human Neuroimaging (University College London). She works on the intersection between philosophy of cognition and computational neuroscience. More precisely, Hipólito applies tools of conceptual modelling to answer philosophical questions of cognition that are compatible with the formalisms of dynamical systems theory. Hipólito has co-edited special issues for *Philosophical Transactions*, *Consciousness and Cognition*, and the Mind and Brain Studies (Springer). She has published work in edited books (Routledge, CUP) and journals (*Australasian Philosophical Review*, *Physics of Life Reviews*, *Progress in Biophysics and Molecular Biology*, *Synthese*, *Network Neuroscience*). Dr. Hipólito's work has been honored with international prizes and awards, including the Portuguese Ministry for Science and Higher Education; the University of Oxford; the Federation of European Neuroscience Societies; and an award by the British Association for Cognitive Neuroscience.

Part I
Technology and the Metaphysics of Mind

Chapter 2
Emergent Mental Phenomena

Mark H. Bickhard

2.1 Introduction

I argue that mental phenomena are emergent in biological phenomena, and, potentially, in artificial systems – though not with current technology. Developing this argument requires addressing several foundational issues. In particular, accounting for the normativity of mental phenomena requires a model of normative emergence, which, in turn, requires a model of genuine metaphysical emergence, which, so I argue, in turn requires an underlying process metaphysics. I will outline these framework preliminaries, in preparation for the discussion of emergent mental phenomena – with a focus on representing and consciousness – and some implications for the possibilities of artificial minds.

The emergence of normativity, given this background framework of process-emergence, occurs in certain kinds of thermodynamic systems. Within the general model of normative emergence, explicated in terms of normative *function*, I address, in turn, representational normativity, basic or primary consciousness, and reflective consciousness. Several conclusions follow from this overall model regarding the possibility of artificial mental systems, and for some other notions, such as that of 'uploading' of persons into computational frameworks.

I begin with an argument for process metaphysics – process metaphysics grounds the further levels of the model.

M. H. Bickhard (✉)
Department of Philosophy and Department of Psychology, Lehigh University, Bethlehem, PA, USA
e-mail: mhb0@lehigh.edu

R. W. Clowes et al. (eds.), *The Mind-Technology Problem*, Studies in Brain and Mind 18, https://doi.org/10.1007/978-3-030-72644-7_2

2.2 Preliminaries

2.2.1 Metaphysics and Emergence

2.2.1.1 Why a Process Metaphysics?

There are several reasons for abandoning standard particle (or substance) metaphysics in favor of a process metaphysics.[1] These include conceptual problems, as well as problems with contemporary physics. For an example of a problematic conceptual problem, consider that a pure point particle model results in a world in which nothing ever happens – point particles have a zero probability of hitting each other. Furthermore, there is nothing to move such particles toward or away from each other, or for them to hold together if they were to ever be in other particles' vicinity. A particle model in which particle interactions occur via *fields* can partially resolve these issues,[2] but a field is a process – a field model is already a process model.

Furthermore, if we look to our best contemporary physics – quantum field theory – we find that there are no particles. Instead, there are various excitations of quantum fields (Fraser 2008; Halverson and Clifton 2002; Hobson 2013; Weinberg 1995).[3] Thus, we find fatal conceptual problems for a particle metaphysics, and no support within physics. Quantum field theory, on the other hand, is a clear process model.

2.2.1.2 Process Metaphysics and Emergence

Emergence is at best mysterious, if not impossible, within a substance or entity metaphysics. How can a new substance or entity 'emerge' from (some organization of) already existing substances or entities?[4] On the other hand, it is fairly easy to see how a new property could be instantiated in a new organization: an organization *is* a property of what is so organized. But, even if such a "new" property were dynamic (not just static), it is difficult to see how it could be a causally efficacious constituent of the world. One of Kim's arguments against emergence makes this clear.

---

[1] Bickhard 2009.

[2] Fields can introduce attractive and repulsive forces.

[3] These excitations and their interactions are quantized in the sense of involving whole integer or half integer amounts. This quantization is the last remnant of 'particle' intuitions, but it is akin to the whole or half integer wavelengths in the vibrations of a guitar string (or, in the case of half integer, perhaps a rope that is free at one end) – and there are no guitar sound (or rope) particles either.

[4] Parmenides argued against change, including emergence, and Democritean atoms and Empedoclean substances were proposed as satisfying this prohibition of fundamental metaphysical change – they can reconfigure and remix, but they do not themselves change. Aristotle accepted this prohibition of fundamental change, and the presupposition has dominated Western thought since then (Gill 1989; Campbell 2015; Bickhard 2009).

In the 'causal regularity' argument, Kim points out that a new configuration of particles might reliably yield a regularity of consequences, including causal consequences, but any such causal consequences would be due solely to the interactions among the particles that instantiated the configuration in the first place (Kim 1991). Configurations can yield causal regularities, but not causal powers.

This argument reveals rather clearly that the underlying assumption is that only particles can bear causal power. Configuration is not a particle, or an entity, or a substance, so it is not even a candidate for having causal power.[5] Configuration is just initial conditions, or boundary conditions, for the real causal interactions among the particles – configuration is just stage setting for the real causal dance. But emergence is supposed to be emergence from or within configuration (organization), so, in this view, it is ruled out *by assumption* that there could possibly be any new causal power emergent in organization. Note that this formally begs the question concerning emergent causal power: the very possibility is assumed to not exist.

In a process metaphysics, however, this default assumption about the locus of causality is flipped: processes are intrinsically organized, and processes influence other processes in crucial part due to their organization. Organization *cannot be delegitimated* as a potential locus of 'causal' influence without eliminating 'causality' entirely – without eliminating 'causality' from the world.[6] But, if organization can manifest 'causal' influence ('power'), then 'new' organization can manifest 'new' – emergent – efficacious influence in the world.

2.2.2 Normative Emergence, Function, Representation

2.2.2.1 Normative Emergence

If emergence per se is a metaphysical possibility, then perhaps the emergence of *normativity* is possible. There are long standing reasons why this should *not* be possible, the first of which is 'simply' that emergence itself is rather difficult to make any sense of within a classic particle or substance model. In addition to that problem, there is also the point that the substance/particle metaphysical world simply has no place, emergent or not, for normativity: it is a metaphysics of fact and cause. This split is enshrined in contemporary thought via (among other bases) Hume's 'argument' against being able to derive norms from facts.

But Hume's 'argument' is (arguably) unsound, and a shift to a process metaphysics not only makes emergence more generally possible, but also opens the possibility of accounting for normative emergence.

[5] For analyses of Kim's more well known argument – the pre-emption argument – showing that it too depends on the same underlying particle assumptions, see (Bickhard 2009, 2015).

[6] The scare quotes are because this kind of ongoing (coupling constant) influence among quantum fields (for example) does not fit well with standard causal chain models of causality.

2.2.2.2 Hume's Argument

Hume did not detail an argument for his "no ought from is" maxim, but claimed that it "seems altogether inconceivable, how this new [normative] relation can be a deduction from others, which are entirely different from it" (see Hume 1978, Book III. Part I. Section I. 469–470). Hume's unstated 'argument' has generated considerable work attempting to explicate and formalize it (e.g., Schurz 1997); I offer an explication of Hume's point that shows it to be valid but unsound.[7] In being unsound, Hume's maxim ceases to be a barrier to the possibility of a model of normative emergence.

A central aspect of deduction is that of definition; e.g., in deducing theorems about triangles from Euclid's axioms, a definition of 'triangle' based on the terms in the axioms (e.g., point and line) is required. Hume's maxim is readily derived if we consider how any new (e.g., normative) terms could be validly introduced in deductions from strictly factual premises. Any new terms must be defined making use of terms already available. These may include those introduced by prior definitions, and those may make use of still prior definitions, but all such definitions must ultimately be in terms that are originally available in the premises. But, given such a hierarchy of definitions, each 'new' term can be back-translated via its definition into prior terms,[8] again through the layers of the hierarchy, till all terms in the conclusion are unpacked into terms in the premises. But, by assumption, those premise terms are all factual, so any valid conclusions must also be strictly factual – not normative. And we have Hume's 'inconceivability' of deduction.

This argument is valid, but it is based on the unstated premise that all definitions are 'abbreviatory' – that all definitions can be back-translated through. But that premise is false, and, if so, the argument is unsound. There was no alternative to abbreviatory, back-translation, definition in Hume's time, but in the nineteenth century *implicit definition* was introduced, with Hilbert being one of its major proponents.[9]

The intuition of implicit definition is that a system of relations – an axiom system for geometry for example – implicitly defines the class of all of its models. It implicitly defines the class of all of the ways in which it can be interpreted that honor all of the relations. Two Xs determine a Y, for example, can be interpreted as "two points determine a line", but it also turns out that it can be interpreted as "two lines determine a point" (the intersection of the lines, so long as intersections at infinity are considered). Implicit definition can also be understood non-formally (Hale and Wright 2000) and also dynamically.[10]

---

[7] As well as that it is related to more general anti-emergentist arguments. See Bickhard (2009, 2015).

[8] I.e., substitute the defining term, phrase, or clause for the defined term. The defined terms are abbreviations of the definiens, so 'fill out' all of the abbreviations.

[9] E.g., Hilbert developed an implicit definitional approach to geometry (Hilbert 1971). See Chang and Keisler (1990) for a modern formal development of implicit definition.

[10] The model of functional presupposition, developed later in the text, is an example of dynamic implicit definition.

The key point for current purposes is that implicit definition is a powerful form of definition (e.g., model theory is based on it) and it is not abbreviatory – it does not permit back-translation. The (re-constructed) Hume argument, therefore, cannot go through, and the "no ought from is" maxim does not necessarily hold. The overall argument is unsound: it involves a false premise concerning definition.

The *barrier* of the Humean maxim is, thus, cleared. But that does not constitute a *model* of normative emergence. I turn to that now.

2.2.2.3 Normative Function

To this point, the discussion has been 'brush clearing' – clearing apparent barriers to emergence and to normative emergence. I now turn to a model of the emergence of normativity in the form of normative function – the sense of function in which it makes sense to distinguish function from dysfunction, as in "This kidney is dysfunctional."

The model of function is grounded on a crucial asymmetry between thermodynamically differing kinds of processes – in particular, between process organizations that are stable in energy wells and those that are (relatively) stable in far from thermodynamic equilibrium conditions. Processes are always ongoingly changing, but some *organizations* of process can remain stable for some time as organizations, and that stability can be of (at least) two differing kinds.

Thus, there are two kinds of such stability that will be of concern here: The first kind are process organizations that remain stable because they are in an 'energy well'. Such organizations will remain stable unless and until some above threshold amount of energy impinges on them that is sufficient to disorganize them – to knock them out of the energy well.

An atom would be a canonical example. It is a furious process of quantum fields that can remain stable for cosmological time periods, if not disrupted. One crucial feature of energy well stabilities is that, if they are isolated, they go to thermodynamic equilibrium and remain in their organization indefinitely.

In contrast are far from equilibrium stabilities of process. Like energy well stabilities, far from equilibrium organizational stabilities can persist for some time. Unlike energy well stabilities, however, they cannot be isolated: being far from equilibrium is a *relational* condition that must be *maintained*. If isolated, they go to equilibrium and the process organization ceases to exist.

A canonical example of this would be a candle flame. If isolated, the flame ceases: if isolated, the process goes to equilibrium and is, thus, no longer far from equilibrium. Far from equilibrium conditions must be maintained.

A candle flame also illustrates a further property: it makes contributions to its own stability. The flame maintains above combustion threshold temperature, vaporizes wax in the wick, melts wax in the candle, and induces convection, which brings in oxygen and removes waste products. It contributes to its own stability in several ways; it is *self-maintaining*.

There is an additional property that a candle flame does not have, but a bacterium does. If a candle flame is running out of candle, it cannot change the process in any way to adapt. A bacterium, in contrast, will tend to swim upward in a sugar gradient, which is a contribution to its stability, but, if it is going toward lower sugar concentrations in the gradient, it will tend to tumble, and then resume swimming.[11] Swimming contributes to stability in some conditions (up a gradient) and not in others, and the bacterium can adjust what it is doing in order to maintain the property of contributing to its own persistence. It self maintains its condition of being self maintenant: it is *recursively self maintaining*.

The crucial point here is that contributions to the maintenance of a far from equilibrium process are *contributory* – they are useful, *functional*, *for* and *relative to* the persistence of that organization of process. This is a normative relationship: it can hold or not hold, and it makes a difference to the system whether it holds or not.[12]

The structure of the model of emergent function, thus, is that:

(1) The asymmetry between energy well and far from equilibrium processes yields
(2) An asymmetry between contributions to the thermodynamic maintenance of far from equilibrium processes and impairments to that maintenance, which, in turn, grounds
(3) The emergent asymmetry between functional and dysfunctional.

The further properties of being self-maintenant and recursively self-maintenant, in turn, begin a hierarchy[13] of more complex forms of autonomy of far from equilibrium systems. Note that this sense of autonomy focuses on the interdependence between a system and its environment – the ability of the system to adjust itself *and* its environment toward functionality for the system – rather than autonomy in the sense of independence or freedom *from* the environment.

2.2.2.4 Representational Truth Value

The normativity of function grounds a further emergent normativity – the normativity of representational truth value. This emergence occurs with respect to a particular function that is necessary for any agent interacting with its world: the function of being able to select what to do next, or to guide ongoing interaction, on the basis (among other things) of what the possible interactions might be in the current situation. That is, there must be some (functional) indications of, anticipations of, further courses of possible interaction in the current situation among which the system (organism) can select.

---

[11] See Campbell (1974). Bacteria can be more complex than this, but the simple example illustrates the point.

[12] The sense in which this is a functionality relative to a system can be illustrated with the case of the beating heart of a parasite, which is functional for the parasite but dysfunctional for the host.

[13] Or, more complexly, a lattice or weave.

A frog, for example, might have indications that it could flick its tongue in any of several directions and eat. An external observer might see a couple of flies and a worm in those directions. Such indications are anticipatory that, if selected, the interaction would proceed as indicated. But such anticipations can be in error: they can be false. Representational truth value is emergent in such anticipations. Truth value bearing anticipatory indication, in turn, is the basis for representing in general, including more complex representing.[14]

2.2.2.5 Content

Furthermore, an indication of the potentiality of an interaction *presupposes* that sufficient supporting conditions for that interaction to succeed hold in the environment – such as, perhaps, a fly or worm. If the anticipatory indication fails, then those supporting conditions did not hold. This implicit presupposition of supporting conditions in the environment, thus, is *about* that environment. Presupposed supporting conditions constitute a model of *content* – the supporting conditions are what are *supposed* to exist in order for the anticipation to hold. Content in this sense, however, is implicit, not explicit (Bickhard 2009).

2.2.2.6 Complex Representing

The model of representing in terms of anticipations of potential interacting has two important resources for modeling more complex representing. The first has already been indicated: the frog has *branching* indications, hopefully triggered by, for example, flies or worms.[15] The second resource is that such indications can iterate. Again, the frog can provide an example: perhaps among the frog's indications are that, if it were to move a little to the left, another pair of worms would come into range. So, indications may be that, some interaction is possible, and, if it were engaged in, it would yield the conditions for further interactions.

Branching and iterating indications of interactive potentialities can link to create potentially complex webs of interactive anticipation. In humans, these webs are vast. I have dubbed such webs as *situation knowledge* – interactive knowledge of what the organism could do in a broad (branched and iterated) sense.

---

[14] This model involves a shift in what is taken to be most centrally criterial for representing. Standard models assume that the crucial property of representing is that of having some sort of denotational or referential relationship with the environment – some sort of critical contact with or correspondence with the environment. In the model outlined above, the criterial property for representing is that of bearing (potential) truth value. Contact with the environment is also centrally important, but, contrary to standard assumptions, contact per se does not yield truth value (Bickhard 2009; Oguz and Bickhard 2018).

[15] "Hopefully" because, for example, they might also be triggered by (contact with) a tossed pebble, in which case they would be false.

Situation knowledge is not constant. It is ongoingly changing on the basis of processes occurring in the environment and on the basis of activities of the organism. Situation knowledge is undergoing constant maintenance and updating. Such processes of keeping situation knowledge up to date is *apperception*.

The situation knowledge web, in turn, is a realm in which more canonical forms of representing can be modeled. For example, a child's toy block will support a (sub-)web of possible interactions that is *closed* in the sense that any manipulatory or perceptual interaction with the block can connect with, lead to, any other such interaction with the block via some intermediary interactions, such as rotations of the block. Further, this internally reachable subweb is itself invariant under many other kinds of processes, such as throwing the block, leaving it on the floor, putting it away in the toy box, and so on. It is *not* invariant under all interactions, however: burning the block, for example, eliminates the support for that situation knowledge subweb. This model of representing a small manipulable object is basically Piaget's model of representing a small object stated within the terms of this model (Piaget 1954). The Piaget model can be borrowed from in this manner because both are 'pragmatic' models, based on action and interaction rather than on correspondence.[16]

2.3 Consciousness: Primary and Reflective

The model of representing supports a model of consciousness with two aspects, primary and reflective. These aspects do not have to occur together, though reflective cannot occur without primary – primary consciousness is what reflective consciousness can reflect upon. But primary forms of consciousness can and do occur without reflective, e.g., in some species and in neonate humans. If this model is correct, consciousness is not a unitary kind of phenomena.

2.3.1 Primary Consciousness

In particular, the model of interaction, situation knowledge, and apperception already outlined provides an account of a process flow that is intrinsically contentful – the contents of situation knowledge, the apperceptive processes that maintain it, and the anticipatory processes that make use of it. This is a flow of content in the sense of anticipations of possibility and relationships among them, not in terms of encoding correspondences (or denotations). It is an intentional flow in terms of the differentiations of the world induced by interactions with that world, and the

[16] This model borrows from Piaget for other phenomena as well, though usually with modifications: Piaget evidences what I take to be errors in some aspects and parts of his model (Bickhard 1988; Bickhard and Campbell 1989).

anticipatory indicative relations among them. It is a partition epistemology (Levine 2009) – partitions induced by differentiations – rather than a correspondence epistemology.[17]

Furthermore, this flow is intrinsically embodied – a body is necessary for *inter*-action. Consequently, it is also situated and from a point of view. The model, thus, accounts for multiple properties of consciousness, properties that arguably exist in simple organisms as well as human beings.[18]

But this does not model all properties of human consciousness (and perhaps some other higher primates). In particular, it does not model conscious reflection, or *reflective consciousness*.

2.3.2 *Reflective Consciousness*

Primary consciousness involves a taking into consideration the agent's relationships with the world and with the potentialities of that world. It is a kind of *awareness* of the world. But primary consciousness does not offer an account of awareness of awareness, of *reflective consciousness*. Interacting is asymmetric; it is normatively 'about'; it involves a normative agent and a world. In particular, interacting, thus awareness, is not in itself reflective.

But reflection can clearly occur: any reflection on the issue is an instance of the phenomenon. The interactivist model argues for a level of *awareness of* processes of *awareness* (interaction, situation knowledge, apperception) that has evolved (in various forms and degrees) in the CNS of some species (Bickhard 2015a, b).

Reflective consciousness, thus, is reflective awareness of processes of awareness. It involves differentiated aspects of the CNS; it requires primary consciousness to reflect upon; but primary consciousness can exist in various species and individuals without the possibility of reflection: consciousness is not a unitary kind of phenomena (Bickhard 2005). Reflection is constituted as a second level interacting with the first level, which, in turn, interacts with the environment.[19]

[17] In general, cognitively simpler organisms will involve less complex situation knowledge webs and more general differentiations – e.g., the differentiation of a "keep swimming" condition.

[18] For consideration of phenomena such as emotion, motivation, and other psychological phenomena, see (Bickhard 2000, 2003).

[19] Note the partial convergence with HOT theories of consciousness (Lau and Rosenthal 2011; Rosenthal 2010).

2.3.3 Experiencing

Primary conscious flow is a strong candidate for modeling *experiencing*. It involves experiencing qualities or properties of interacting with the world. Reflective consciousness, in turn, involves experiencing the properties or qualities of primary experiencing – experiencing experiencing.

These properties of the experiencing of experiencing are commonly reified into discrete elements called qualia. This is incorrect: experiencing is a flow. But, worse, qualia are often taken to not only be (discretized) experiencing of experiencing, but also to be *constitutive* of *basic* experiencing – as in sense data models. This would make qualia the experiencing of qualia. This is a tight metaphysical circle that makes understanding experiencing ultimately impossible (Bickhard 2005).

Qualia (overlooking the assumed discreteness) are *results of reflection*, not constituents of what is reflected on.[20] Reflective experiencing is much easier to understand if this is taken into account. Furthermore, what are at times taken to be the "easy" problems – e.g., normative representing – are much harder to understand than often considered – they involve, for example, the emergence of normativity. The realm of consciousness and experience looks significantly different when viewed in these interactive terms.

2.3.4 An Argument Against the Possibility of the Emergence of Conscious Experience

There are strong arguments in the literature against any such possibility of the emergence of consciousness and experiencing. As with the problems induced by particle metaphysics and by Hume's "argument", I will address one class of these arguments with an intent of brush clearing (again) – showing that the arguments are unsound, in that they make an underlying false assumption.

The argument that I wish to address has the following general form: We can model numerous "easy" problems, such as representation, in terms of causal functionalism – in terms of, for example, symbolic or connectionist encodings. But causal function is 'just' a standard causal relation that is picked out of a realm of causal consequences as being relevant to some consideration, such as constituting part of what makes a computer. A transistor, for example, has multiple causal consequences, such as creating heat, but the only one that counts as causally functional is the (perhaps) switching function that it introduces in a circuit.

But causality is indifferent to consciousness and experience. A causally identical system or organism could have wildly different experiences, or none at all (a zombie). Possibilities of inverted qualia, dancing qualia, and so on are perfectly consistent with whatever causal functional processes make up a person, so that experiential

[20] This is basically Dewey's criticism of Russell's sense data model (Dewey 1960; Tiles 1990).

realm is independent of those causal functional processes: the experiencing could be wildly different, or not exist at all.

Causal functionality seems adequate to many 'mental' phenomena – those are 'easy' problems – but cannot be adequate to the qualia of experience – that is the 'hard' problem (e.g., Chalmers 1996).

One crucial assumption in this argument is manifest: every functional relationship is 'just' a selected causal relationship. This ignores, for example, the possibility of normative function. It might be argued, however, that even normative function is still 'just' cause, just cause selected by some sort of *intrinsic* functionality, so it is still intrinsically 'just' cause and therefore cannot account for the qualia of experience.

That line of dispute can be continued further, but there is a deeper problem that I wish to point out that undercuts that framework of issues: there is an assumption in this argument, including in its reliance on 'cause', that all (crucial) relationships are *external*.

The distinction between metaphysical internal and external relationships is mostly lost in contemporary philosophy. It was important at the turn into the twentieth century, and for decades after, but Russell tried to reject internal relations and Quine pretty successfully did so.

The distinction is between relations that are in some basic sense essential to something's being what it is (internal) and some relations that are irrelevant to what something is. An external relationship might be between an effect and its cause: it would be that effect even if from a quite different cause. An example of an internal relationship might be that between an arc of a circle and the point that is the center of that circle: it could not be an arc of that circle if it did not have that relation to that point.

Note that a background assumption of a particle metaphysics in which particles are independent particulars is a framework in which the basic level of existence is composed of entities that cannot have internal relations (if they did have internal relations, they would not be independent particulars, thus not particles; Seibt 1996, 2009, 2010, 2012; Campbell 2015).

Similarly, a mechanistic, causal functional framework, is one in which all relations are external (cause is a classical example of an external relation) *by assumption*. The case of the presuppositions of normatively functional anticipations, however, is a kind of *internal* relation: in no universe with thermodynamics like this one could there exist normative anticipations of possible interactions that did not have (was not related to) presuppositions of the possible supports for those anticipations.

Even if this model of presupposition is not accepted, the background assumption in the "hard" problem arguments of external relations ignores the very possibility of internal relations, and, thus, the arguments are unsound: if the relations between system processes and conscious experience are internal, then zombies, inverted and dancing qualia, and so on, are impossible. So the assumption that the relations are external has to be established in order for these arguments to go through, but that

assumption is not even addressed. That suffices for the 'brush clearing' of these arguments: they are not a barrier because they are unsound.

2.4 Artificial?

The model outlined is *not* a computational model.[21] As such, it significantly alters the questions concerning the possibility of artificial intelligence, agency, cognition, and consciousness: If this model, or anything like it, is correct, then none of these phenomena can be constituted as strictly computational systems.

The model, however, does not preclude the possibility of artificial systems that have emergent mental properties. But this cannot be done with contemporary (mechanistic/computational) technology. Instead, the emergence of normativity, including in its myriad particular instances, requires particular kinds of far from thermodynamic equilibrium systems, not (just) mechanistic systems (Bickhard 2007).

And, though not impossible in principle, creating artificial complex recursively self-maintenant systems requires creating something that constitutes life (at least metabolically). A major challenge.

2.4.1 Uploading?

One consequence of the shift in this model to a continuous dynamic framework is the impossibility of what has been called "uploading" – the uploading of a person into a computational condition.

Computation requires special conditions to be imposed and maintained on underlying continuous processes. In particular, the physical differences that underlie what are called ones and zeros must be maintained as detectably distinct and must be converted consistently across all physical substrates involved in the computational processes: that is, magnetization up and down must be kept distinct, and must be converted consistently across categories of high voltage/low voltage, high electromagnetic pulse/low pulse, high charge/low charge, spin up/spin down, light pulse/no light pulse, and so on. Maintaining such differences and the consistencies of conversions among them is required in order for computation on such categories (e.g., ones and zeros) to be possible, but such maintenance is in the face of physical dynamic tendencies toward smearing and erasure of the distinctions. Computers are good at this; otherwise they wouldn't be consistent, or would fail altogether.

[21] It is, in fact, *anti*-computational. Furthermore, for similar reasons of fundamental incoherence in models of representing (as well as other phenomena), it is also anti-connectionist – not with respect to the technologies per se, but with respect to claims that they might capture the nature of representing. For discussions of this and some other models and positions in the literature, see (Bickhard 2009, 2014, 2016a, b; Bickhard and Terveen 1995).

The very notion of uploading, however, depends on a background assumption of such discrete causal functionalism. If the model outlined above is correct, then there are two fatal problems with this assumption: (1) *causal* functionalism does not suffice: causal functionalism is a mechanistic model and cannot realize the necessary normativities, and (2) real dynamic phenomena cannot be captured by discrete approximations. For example, chaotic processes can be useful to an organism for the sake of unpredictability-in-principle in situations of competition or predator/prey interactions – it can be unadaptive to be predictable, and chaos makes predictability impossible. This is because chaotic processes are sensitive to close to infinitesimal differences in (initial and/or boundary) state and no process can measure such fine differences, so no reliable prediction can be made. So, insofar as anything like chaos is important in the realization of mental processes, no possible measurement is capable of capturing the crucial differences in state. So, no uploading can occur.

Furthermore, emergence is important, according to this model, not only in the broad sense of normative, functional, representing, and other kinds of emergence, but also in ongoing processes, such as emergent variations in mental variation and selection problem solving processes (Bickhard 2002), and, again, this cannot be captured in discrete measurements or discrete systems: (1) it is likely to itself be chaotic, and (2) the self-organizational properties of irreversible processes is involved (Bickhard, 2002), and this too involves fine sensitivity to initial and boundary conditions.

The very notion of uploading, thus, depends on the background assumption of the adequacy of computational models to mental phenomena. The interactivist model precludes that adequacy, and thus precludes the possibility of uploading.

2.5 Conclusion

Artificial systems with emergent mental properties are possible in principle, but not with current technology. Appearances that current information processing – computational, connectionist, and so on – technology might be adequate are false. What is required are particular kinds of far from thermodynamic equilibrium processes that go beyond 'mechanistic' causal functional models and technologies.

I have outlined a model of the emergence of normative phenomena, particularly normative function, representing, and experiencing in this chapter, and investigated some of the consequences for the possibility of artificial experiencers. The very possibility of such emergences, in turn, depends on a process metaphysics, which enables a model of metaphysical emergence, which enables a model of normative emergence.

Thus, the framework of process, emergence, normative emergence, and the specific models of various kinds of normative emergence form an integrated conceptual whole. No one part of it stands alone. Within this framework, the creation of artificial experiencing systems is possible, but only with a technology that can design

and create complex recursively self-maintaining far from thermodynamic equilibrium systems.

References

Bickhard, M. H. (1988). Piaget on variation and selection models: Structuralism, logical necessity, and interactivism. *Human Development, 31*, 274–312.

Bickhard, M. H. (2000). Motivation and emotion: An interactive process model. In R. D. Ellis & N. Newton (Eds.), *The caldron of consciousness* (pp. 161–178). Amsterdam: J. Benjamins.

Bickhard, M. H. (2002). Critical principles: On the negative side of rationality. *New Ideas in Psychology, 20*, 1–34.

Bickhard, M. H. (2003). An integration of motivation and cognition. In Smith, L., Rogers, C., Tomlinson, P. (Eds.) *Development and motivation: Joint perspectives* (Monograph series II, pp. 41–56). Leicester: British Psychological Society.

Bickhard, M. H. (2005). Consciousness and reflective consciousness. *Philosophical Psychology, 18*(2), 205–218.

Bickhard, M. H. (2007). Mechanism is not enough. In Q. Gonzalez, M. Eunice, W. F. G. Haselager, I. E. Dror (Eds.) Mechanicism and autonomy: What can robotics teach us about human cognition and action? (Special issue of pragmatics and cognition, Vol. 15, issue 3, pp. 573–585). Amsterdam: Benjamins

Bickhard, M. H. (2009). The interactivist model. *Synthese, 166*(3), 547–591.

Bickhard, M. H. (2014). What could cognition be, if not computation … or connectionism, or dynamic systems? *Journal of Theoretical and Philosophical Psychology, 35*(1), 53–66.

Bickhard, M. H. (2015). The metaphysics of emergence. *Kairos, 12*, 7–25.

Bickhard, M. H. (2015a). Toward a model of functional brain processes I: Central nervous system functional micro-architecture. *Axiomathes, 25*(3), 217–238.

Bickhard, M. H. (2015b). Toward a model of functional brain processes II: Central nervous system functional macro-architecture. *Axiomathes, 25*(4), 377–407.

Bickhard, M. H. (2016a). The anticipatory brain: Two approaches. In V. C. Müller (Ed.), *Fundamental issues of artificial intelligence* (pp. 259–281). Cham: Springer.

Bickhard, M. H. (2016b). Inter- and En- activism: Some thoughts and comparisons. *New Ideas in Psychology, 41*, 23–32.

Bickhard, M. H., & Campbell, R. L. (1989). Interactivism and genetic epistemology. *Archives de Psychologie, 57*, 99–121.

Bickhard, M. H., & Terveen, L. (1995). *Foundational issues in artificial intelligence and cognitive science: Impasse and solution*. Amsterdam: Elsevier Scientific.

Campbell, D. T. (1974). Evolutionary epistemology. In P. A. Schilpp (Ed.), *The philosophy of Karl Popper* (pp. 413–463). LaSalle: Open Court.

Campbell, R. (2015). *The metaphysics of emergence*. New York: Palgrave Macmillan.

Chalmers, D. J. (1996). *The conscious mind*. Oxford: Oxford University Press.

Chang, C. C., & Keisler, H. J. (1990). *Model theory*. North Holland.

Dewey, J. (1960/1929). *The quest for certainty*. New York: Capricorn Books.

Fraser, D. (2008). The fate of "particles" in quantum field theories with interactions. *Studies in History and Philosophy of Modern Physics, 39*, 841–859.

Gill, M.-L. (1989). *Aristotle on substance*. Princeton: Princeton University Press.

Hale, B., & Wright, C. (2000). Implicit definition and the a priori. In P. Boghossian & C. Peacocke (Eds.), *New essays on the a priori* (pp. 286–319). Oxford: Oxford University Press.

Halvorson, H., & Clifton, R. (2002). No place for particles in relativistic quantum theories? *Philosophy of Science, 69*(1), 1–28.

Hilbert, D. (1971). *The foundations of geometry*. La Salle: Open Court.

Hobson, A. (2013). There are no particles, there are only fields. *American Journal of Physics, 81*, 211. https://doi.org/10.1119/1.4789885.
Hume, D. (1978). *A treatise of human nature*. Index by L. A. Selby-Bigge; notes by P. H. Nidditch. Oxford: Oxford University Press.
Kim, J. (1991). Epiphenomenal and supervenient causation. In D. M. Rosenthal (Ed.), *The nature of mind* (pp. 257–265). Oxford: Oxford University press.
Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences, 15*(8), 365–373.
Levine, A. (2009). Partition epistemology and arguments from analogy. *Synthese, 166*(3), 593–600.
Oguz, E., & Bickhard, M. H. (2018). Representing is something that we do, not a structure that we "use": Reply to Gładziejewski. *New Ideas in Psychology, 49*, 27–37.
Piaget, J. (1954). *The construction of reality in the child*. New York: Basic.
Rosenthal, D. (2010). How to think about mental qualities. *Philosophical Issues, 20*, 368–393.
Schurz, G. (1997). *The is-ought problem: An investigation in philosophical logic* (Trends in logic, Vol. 1). Dordrecht: Kluwer Academic.
Seibt, J. (1996). The myth of substance and the fallacy of misplaced concreteness. *Acta Analytica, 15*, 119–139.
Seibt, J. (2009). Forms of emergent interaction in general process theory. *Synthese, 166*(3), 479–512.
Seibt, J. (2010). Particulars. In R. Poli & J. Seibt (Eds.), *Theories and applications of ontology: Philosophical perspectives* (Vol. 1, pp. 23–57). New York: Springer.
Seibt, J. (2012). Process philosophy. In *Stanford encyclopedia of philosophy.* http://plato.stanford.edu/entries/process-philosophy/
Tiles, J. E. (1990). *Dewey*. London: Routledge.
Weinberg, S. (1995). *The quantum theory of fields. Vol. 1. Foundations*. Cambridge: Cambridge University Press.

Mark H. Bickhard is the Henry R. Luce Professor in Cognitive Robotics and the Philosophy of Knowledge at Lehigh University. He is affiliated with the Departments of Philosophy and Psychology, and is director of the Institute for Interactivist Studies. His work ranges from process metaphysics and emergence to consciousness, cognition, language, and functional models of brain processes to persons and social ontologies.

Chapter 3
Technology and the Human Minds

Keith Frankish

3.1 Introduction

Over the last 40 years, many psychologists have come to adopt some form of *dual-process* theory. Such theories hold that human cognition is supported by two distinct types of processing – a fast, automatic, unconscious type, and a slower, controlled, conscious one – which can yield different and sometimes conflicting results. The distinction corresponds to the everyday one between intuition and reflection, the former delivering spontaneous responses that just feel right, the latter more considered responses for which one can give some explicit justification. These types of processing are sometimes said to be associated with two brain systems, System 1 and System 2, the first evolutionarily ancient and largely shared with other animals, the latter more recent and distinctively human. Some dual-process theorists speak of our having two *minds*, an intuitive *old mind* (System 1) and a reflective *new mind* (System 2) (Evans 2010).

Such views have obvious implications for cognitive enhancement and artificial intelligence (AI). If we do have something like two minds, then the projects of enhancing and replicating human intelligence will each also assume a dual aspect. Dual-process views face some problems, however, and it is hard to see how System 2 could be modelled artificially. These problems, I believe, dictate a reinterpretation of dual-process theory, which pictures the two minds as levels of organization rather than distinct systems. The new mind should be seen, not as a brain system, but as a virtual one, formed by culturally transmitted habits which restructure the activities of the old mind. This reinterpretation helps to resolve some of the problems for

K. Frankish (✉)
Department of Philosophy, The University of Sheffield, Sheffield, United Kingdom

Open University, Milton Keynes, UK

Brain and Mind Programme, University of Crete, Crete, Greece
e-mail: k.frankish@sheffield.ac.uk

R. W. Clowes et al. (eds.), *The Mind-Technology Problem*, Studies in Brain and Mind 18, https://doi.org/10.1007/978-3-030-72644-7_3

dual-process theory and makes the project of artificially creating a System 2 mind somewhat more tractable.

In this chapter I shall explore these issues, explaining the re-interpretation of dual-process theory, looking at its implications for projects of cognitive enhancement and AI, and assessing the risks of those projects as they appear in this new light. I begin, however, by introducing the dual-process approach.

3.2 Dual Processes

Dual-process and dual-system theories grew out of experimental work in cognitive and social psychology from the 1970s onwards (for on overview, see Frankish 2010). The theories were formulated in a series of important papers and books published in the late 1980s and 1990s (e.g., Chaiken and Trope 1999; Chen and Chaiken 1999; Epstein 1994; Evans 1989; Evans and Over 1996; Petty and Cacioppo 1986; Sloman 1996; Stanovich 1999; Stanovich and West 2000) and brought to a wider audience in several books published over the next decade or so (Evans 2010; Kahneman 2011; Stanovich 2004).

Many variants of the dual-process approach have been developed, differing in detail but agreeing on the fundamentals. A composite account incorporating the most common claims runs as follows. There are two types of processing ('thinking') involved in human reasoning, decision making, and social cognition: Type 1 and Type 2. Type 1 processing is typically fast, automatic, effortless, non-conscious, associative, parallel, high-capacity, and undemanding of working memory. It is highly contextualized, draws on implicit knowledge acquired from past experience, and delivers responses that may be adaptive in real-world settings but often deviate from rational norms, manifesting cognitive biases, stereotype effects, and emotional influences. Type 2 processing, by contrast, is typically slow, controlled, effortful, conscious, rule-governed, serial, low capacity, and demanding of working memory. It is more abstract, draws on explicit knowledge and learned rules of inference, and is more likely to deliver responses in line with normative principles. Type 2 processing is also linked to hypothetical thinking – evaluating candidate actions in imagination and simulating alternative perspectives and scenarios. For this, we must entertain 'secondary' representations, which are decoupled from the world and do not directly affect behaviour, and this is held to require Type 2 processing. Finally, the propensity to use Type 2 processing shows high individual variability and is correlated with measures of general intelligence.

Dual-process theorists differ as to how the two processes are related. Some see them as operating independently and competing for control of behavior. Others adopt a *default-interventionist* model, according to which Type 1 processes supply rapid default responses, which can be intervened upon and overridden by Type 2 processes. On this view, Type 1 processes are also responsible for triggering Type 2 processing and for selecting information for it to use (Evans 2006; Kahneman 2011).

Dual-*system* theories propose a broader architectural basis for the two types of processing, which assigns them to different mental systems, System 1 and System 2. System 1 is taken to be composed of multiple subsystems, many evolutionarily ancient, all of which operate in a Type 1 way (e.g., Stanovich 2004). These include perceptual, motivational, and emotional systems, learning and conceptual systems (perhaps specialized for particular tasks, such as navigation, foraging, social cognition, theory of mind, and language), and procedures for learned skills practiced to automaticity, such as reading and driving.[1] System 2, on the other hand, is thought of as a single, low-capacity system which can manipulate explicit representations in working memory. It is flexible, responsive to instructions, and uniquely human.

There is a mass of evidence for the dual-process picture, from three independent sources (for a summary and illustrative references, see Evans and Stanovich 2013). First, there is evidence from response patterns in reasoning and decision-making tasks. Typically, participants give one of two answers, the first intuitively plausible but normatively incorrect, the second less obvious but correct, and experimental manipulations can influence which it is. For example, time pressure leads to increased production of the intuitive answer (rather than random responding), whereas clear task instructions promote the normatively correct one. This strongly suggests that two different mechanisms are in play, one fast and intuitive, the other slow and reflective, each of which delivers a specific answer. A second source of evidence comes from work on individual differences. There is a positive correlation between tendency to give the normative responses on reasoning tasks and general intelligence, which is explained on the hypothesis that those of higher general intelligence have a greater capacity to engage in and sustain Type 2 processing and to override intuitive Type 1 responses. (This is not surprising, given that Type 2 processing requires working memory, and working memory capacity itself correlates with general intelligence.) Third, there is neuroscientific evidence. Imaging studies indicate that different neural structures are involved in the production of responses associated with each type of processing, Type 2 responses typically following activation of prefrontal and frontal cortical regions that are not involved in Type 1 responding.

This dual-system view has a common-sense appeal, and something like it has been tacitly acknowledged for centuries (Frankish and Evans 2009). Many early modern philosophers agreed with Descartes in identifying the mind with the conscious mind, understood as an immaterial substance that is the arena of pure thought. But they also recognized that much human and animal behavior occurs without conscious thought and must be supported by complex nonconscious mechanisms of some kind. (Descartes himself fully recognized this; Descartes 1984, p. 161.) The development of scientific psychology in the nineteenth century saw the gradual acceptance that these processes were genuinely mental, involving non-conscious

[1] The basic dual-system framework is compatible with a spectrum of views as to the nature of the evolved components of System 1, from ones which posit multiple domain-specific modules (e.g., Carruthers 2006) to ones which hold that learning is domain-general and that specialized systems are *cognitive gadgets* installed by cultural processes during individual development (Heyes 2018).

perceptions and thoughts, operating independently of the conscious mind. More recently, with the development of the computational theory of mind and modern cognitive science, non-conscious processes increasingly took center-stage in the explanation of human behavior, with the conscious mind sometimes being demoted to the role of a rationalizer (Wegner 2002).

In the history of AI too, we can see implicit acknowledgement of the two-systems distinction. Early AI researchers focused on abstract reasoning and decision making, which they sought to model in computational terms, with the aim of creating artificial general intelligence. Lack of success in this project led many researchers to turn to a bottom-up approach, seeking to create embodied, robotic systems with specific behavioral competences (e.g., Brooks 1991; Steels and Brooks 1995). From a dual-process perspective, this was simply a switch of focus from System 2 to System 1.

3.3 Problems

Despite its attractions, dual-process theory has its critics (e.g., Gigerenzer 2010; Keren and Schul 2009; Kruglanski and Gigerenzer 2011; Melnikoff and Bargh 2018; Osman 2004, 2018). A common objection is that it is highly unlikely that the various features ascribed to each process (fast vs slow, automatic vs controlled, non-conscious vs conscious, etc.) align so neatly, excluding crossover processes that are, for example, fast but controlled. Critics also object to the suggestion that intuitive processing is always biased and reflective processing always normatively rational.

Dual-process theorists respond by clarifying the scope of their claims (e.g., Evans and Stanovich 2013; Pennycook et al. 2018). They explain that the features ascribed to each process are not all defining ones and that the core distinction can be drawn more simply. Evans and Stanovich identify autonomy (lack of attentional control) as the defining characteristic of Type 1 processing, and the use of working memory and support for decoupled representations as those of System 2 (Evans and Stanovich 2013; Stanovich and Toplak 2012). The other features commonly ascribed to each system are held to be merely typical correlates of these defining features. For example, Type 2 processing is typically slow and serial because it loads on working memory, which is a limited resource. As Evans and Stanovich stress, this allows for considerable variation in the mode of Type 2 thinking, since individuals may use many different procedures and strategies for manipulating explicit representations in working memory, reflecting their individual 'thinking dispositions' or 'mindware' (Stanovich 2009a, 2011). For the same reason, it is wrong to think that Type 2 processes always deliver normatively correct responses and that all cognitive errors are due to Type 1 processes. It is true that this pattern is often observed in experimental settings designed to create conflict between the two kinds of processing, but there is no reason to think that it is a universal one. Type 2 processing may often deliver incorrect or biased responses, owing to inattention, misunderstanding, or poor strategy (buggy mindware) (Evans 2006, 2007; Stanovich 2009b).

Conversely, intuitive Type 1 processing may often deliver optimal responses, at least in favorable conditions.

Recently, theorists broadly sympathetic to the dual-process approach have raised more specific worries, especially about the relation between the two systems. These have prompted proposals for the revision or refinement of the framework, though without undermining the case for a qualitative distinction along the general lines proposed (De Neys 2018).

There is, however, another, more general problem I want to raise for the dual-process theory. It concerns Type 2 processing. What exactly is the *mechanism* by which this processing operates? Calling it a *system* implies that the reflective mind is a self-contained device, which takes inputs from System 1 but processes them using its own proprietary mechanisms. Theorists identify various components of this device, including working memory, explicit decoupled representations, and executive control processes, but these do not in themselves amount to a reasoning system. What is the engine that manipulates the explicit representations in working memory, in accordance with rules of inference or other procedures? Dual-system theorists are strangely silent on this.

There is a related evolutionary worry. If System 2 is a self-contained device, it must be an extraordinarily powerful one. We can turn our conscious minds to any problem. We can think about things distant in time and space and about abstractions and hypothetical scenarios. We can construct rational arguments, form and evaluate novel ideas, devise complicated plans of action, and much more. How and why did such a system evolve? Although the human brain is much larger than the brains of other animals, its evolution seems to have involved the addition of new specialist subsystems, such as ones for language, mindreading, and social cognition, and the enhancement of existing ones, rather than the installation of a completely new general-purpose reasoning system (Carruthers 2006). Indeed, it is hard to see what evolutionary pressures there could have been for the development of such a system. Having a capacity for flexible, abstract, rule-governed deliberation is advantageous in the modern world (a world that is largely the creation of our human minds), but it is hard to see why it would have been required in the ancestral environment in which our species evolved. Cognitive flexibility is certainly useful, but to build in general intelligence seems like a massively overengineered solution to any specific environmental challenges our ancestors might have faced.

3.4 Type 2 Thinking as an Activity

I want to suggest a reinterpretation of dual-process theory, which helps to address these problems.[2] The key idea is that some thinking is an intentional activity, something we *do*. The distinctive thing about intentional actions, as opposed to other bodily movements and processes, is that they are under voluntary control, responding to our beliefs and desires. We perform them because we *want* to – either because we enjoy them or because we believe they will further some goal we have. These reasons need not be consciously entertained. Most of our behavior is unreflective: we walk, talk, drive, and go about our daily lives without giving much conscious thought to the reasons for our actions. But the actions are still intentional ones, directed to our goals and guided by our beliefs. The defining feature of Type 2 thinking, I propose, is that it involves performing intentional actions in this sense. Type 1 processing, by contrast, is a wholly automatic process, which occurs without our needing to do anything.

How could reasoning be an intentional activity? Consider solving a long division problem with pencil and paper, following the procedure you were taught at school. We write out the numbers in a certain format, then do a series of simpler calculations, each step building on the previous one, until we arrive at the solution. This involves a series of actions – writing down various numerals in certain locations – which are performed with the goal of solving the problem. But how do we know which actions to perform at each step – which numerals to write and where? What is the reasoning mechanism that takes us from step to step, from one set of symbols to the next? The answer of course is that it is System 1. The answer to each subproblem comes to us intuitively, courtesy of automatic Type 1 processes. When we need to subtract two from seven, say, we just see that the answer is five, and write it down. Each step in the controlled, conscious procedure is driven by intuitive Type 1 processes which are neither controlled nor conscious. Indeed, the role of the procedure is precisely to break down a complex problem that we cannot solve intuitively into smaller problems that we can. The process is, we might say, one of deliberative *mastication*.

A similar process can be used to reason in a more exploratory way. Skilled mathematicians can combine various pen-and-paper procedures, supported by a rich intuitive understanding of the subject, to explore novel theoretical possibilities. Again, the manipulation of written symbols allows them to break down a complex problem into intuitively manageable chunks.[3]

---

[2] This interpretation draws on suggestions by Dennett and Carruthers, among others (Carruthers 2006, 2009; Dennett 1991). For further explorations of the view, see Frankish (1998, 2004, 2009, 2018). It is possible that some dual-process theorists always intended this interpretation and that it is implicit in the characterization of Type 2 processing as *controlled*. If so, then the present proposal is more an explication than a reinterpretation.

[3] The physicist Richard Feynman insisted that his notes were not a record of work done in his head but the very working itself. "No, it's not a record, not really. It's working. You have to work on paper, and this is the paper. Okay?" (quoted in Gleick 1992, p. 409).

How exactly does intentional reasoning like this work? It is useful to think of it as operating by means of what Daniel Dennett calls *autostimulation* (Dennett 1991, Ch. 7). In creating and manipulating external symbols we are cognitively stimulating ourselves, providing new inputs to our Type 1 mental processes. Our perceptual systems detect and interpret the symbols we create, and conceptual, emotional, and motivational systems get to work on the problem of how to respond (what to write next and where). These systems compete for control of motor systems, leading to a further action, which forms the next step in the sequence. We also create drawings and diagrams to help us solve problems and evaluate options. Think of making sketches to experiment with designs for a garden or for the layout of furniture in a room. Again, the process is autostimulatory. We sketch a design, examine it, and our autonomous mental processes generate an evaluative response. Perhaps the design looks ugly or unbalanced or just wrong somehow.

But the most powerful means of autostimulation is speech. By talking to ourselves we can work our way through a tricky problem. We question ourselves ('Where did I leave the remote?'), guide ourselves ('That's the earth pin, so this must be the live'), prompt ourselves ('It begins with a T'), encourage ourselves ('You can do it!'), chide ourselves ('Focus!'), and so on. Again, these utterances are intentional actions, performed with the goal of solving our current problem. They are heard and processed like other utterances and interpreted as requiring some response. Type 1 processes get to work on the task and, with luck, generate a further utterance or other action which either solves our problem or takes us a step closer to a solution. Sometimes we conduct a dialogue with ourselves, posing questions and answering them as a way of thinking through the options. We also create extended arguments, moving from one utterance to another in accordance with simple inferential principles we have been taught or have picked up in the course of debate with others. And as with mathematical reasoning, we can combine a variety of techniques to explore a problem space, using utterances as cognitive stepping stones. Language provides an excellent medium for such flexible, reflective thinking, having an open-ended representational capacity and a syntactic structure that facilitates logical inference.

Intentional reasoning can also be done covertly, in the head. Instead of producing overt symbols, sketches, and utterances, we can create mental images of those things. The claim that we can intentionally create mental imagery is not controversial (just try visualizing your front door or saying your address to yourself in inner speech). In the case of inner speech, this probably involves mentally rehearsing the action of saying the words in question (which generates sensory representations of hearing them), but in other cases it seems to involve the intentional direction of attention in order to stimulate sensory activity associated with relevant stimuli or with episodic memories (Carruthers 2015). In either case, the imagery produced has an autostimulatory effect. Attention sustains the representations in working memory, resulting in their being made available ('globally broadcast') to all Type 1

subsystems, which process them as they would representations generated by external stimuli.[4]

Mental imagery allows the internalization of many external problem-solving activities, in particular those using speech. Processes of self-questioning, self-guiding, self-prompting, argument construction, and inner dialogue can all be conducted silently in one's head.[5] Imagery also allows the development of a wide range of new problem-solving strategies, in which imagined scenarios serve as proxies for aspects of the world. To take an example frequently used in the literature on mental imagery, if you want to know how many windows there are in your house, you can visualize each room in turn and count the windows. Imagery can also be used to evaluate plans and hypotheses before committing to them. If you are trying to decide where to go for a picnic, you can visualize the different candidate locations and see what emotional reactions they evoke. Visual imagery, together with imaged utterances, can thus provide the decoupled 'secondary' representations needed for hypothetical thinking.

This is not the place to attempt a full survey of the various techniques of imagistic autostimulation, but it is safe to say that there are many of them and that they can be flexibly combined in an exploratory way. It is worth stressing that autostimulatory processes needn't be pre-planned. We don't need to know precisely which autostimulations to generate in order to solve a problem. (If we did, then we would in effect already have solved it.) Rather, we follow a process of trial and error and may hit many dead ends before we reach a solution. At the same time, however, the process needn't be completely random. We may have picked up useful tricks and developed hunches about what will work, based on past experience.

Now, my proposal is that the core distinction between Type 1 and Type 2 processing concerns the role of intentional autostimulatory actions. Type 2 processes constitutively involve such actions, whereas Type 1 processes do not. (Since they are not under intentional control, we may continue to speak of Type 1 processes as *autonomous*.) Note that I do not restrict Type 2 processes to ones that occur 'in the head', using sensory imagery. The defining characteristic of Type 2 reasoning is that it involves intentional autostimulatory action. Whether the actions are covert or overt is incidental. Of course, on this view Type 2 processing *also* involves Type 1 processing and is driven by it; but there is still a qualitative difference between the two. Type 1 processes do not involve the performance of intentional actions and are not mediated by perceptions or sensory imagery.

This distinction subsumes the other core distinctions that have been proposed: Intentional autostimulation loads on working memory and supports cognitive decoupling since the perceptual or imagistic representations involved are held in working memory and can represent non-actual states of affairs. It also explains why

[4] For detailed proposals about the neural mechanisms involved in this kind of sensory-based reflective reasoning, see Carruthers 2006, 2015.

[5] Of course, not all intentional reasoning processes can be internalized. When it is necessary to keep referring back to previous steps, as in doing a long division, our working memory capacity is soon exceeded and an external record is required.

Type 2 thinking has the typical correlated features. Autostimulation is conscious because the representations generated are globally broadcast (global broadcast is widely agreed to be sufficient for consciousness in the access sense and at least correlated with consciousness in the phenomenal sense).[6] It is controlled because it is an intentional action, slow and effortful because it requires controlled attention, serial because we can perform only one action at a time, and so on.

3.5 A Virtual Mind

This view of Type 2 thinking has implications for the evolution of the new, 'System 2' mind. This did not require the creation of a new general-purpose reasoning system, or indeed of any completely new neural structures. The engine of Type 2 thinking is provided by the collection of specialist perceptual, conceptual, emotional, and motivational subsystems which constitute the old, System 1 mind, and which evolved in response to specific adaptive pressures.[7] The other key ingredients required for Type 2 thinking were almost certainly already in place too. Forms of working memory, attention, episodic memory, and executive control are found in other animals (Carruthers 2015, Ch. 8), and natural language probably developed initially for social purposes.[8]

This suggests that the development of Type 2 thinking was predominantly a process of cultural evolution, involving the discovery and transmission of habits of autostimulation. It is plausible to see this process as the privatization and then internalization of certain social practices. Humans began by cognitively stimulating each other, helping their peers solve problems by offering suggestions, giving advice, asking questions, making sketches, and so on. They also developed practices of public argumentation, setting out arguments in favor of their ideas and plans. Later, they privatized these habits, providing a similar commentary on their own activities and constructing arguments in private. Finally, they internalized this commentary and developed further self-stimulatory tricks using mental imagery.

There may have been some relatively minor neural adaptations to support the process. Individuals who had discovered the trick of autostimulation would have had a huge advantage over their peers, creating selectional pressure for neural

[6] In fact, I believe that access consciousness is the only kind there is and that phenomenal consciousness is illusory (Frankish 2016). But that is another – though related – story.

[7] Some writers have argued that humans have a specialist argumentation system (of the Type 1 kind), whose function is to produce rational arguments for use in debate with one's peers (Mercier and Sperber 2011). Such a system would obviously be a great asset in supporting Type 2 thinking, helping to generate cogent arguments in inner speech, but it is still a precursor system, which evolved for social purposes.

[8] Speculating about the origins of language is a notoriously risky business, but I think it is safe to assume that its evolution was initially driven by the needs and opportunities of social life, though its co-option for cognitive purposes may have fostered its further development. I assume that the evolutionary process itself was a combined biological and cultural one (Dennett 2017).

adaptations that favored the automatic acquisition and elaboration of the trick – a process known as the Baldwin effect (Dennett 1991). But techniques of intentional reasoning still have to be learned, and a parallel process occurs in child development, as psychologists in the Vygotskyan tradition stress (e.g., Diaz and Berk 1992; Vygotsky 1986; Winsler et al. 2009). Adults *scaffold* children's cognitive development by offering guidance, suggestions, and instructions, which enable children to work through problems they could not have solved on their own. Children then imitate this commentary in self-directed ('private') speech, providing the scaffolding for themselves. Finally, they internalize this private speech as inner speech.

This reinterpretation of dual-process theory casts talk of dual systems in a new light. On this view, there is just one neural system – the collection of 'System 1' subsystems, together with attentional and executive systems and working memory. Note that this claim is compatible with the neuroimaging evidence for dual-*process* theory mentioned earlier. The claim is not that exactly the same subsystems are involved in generating a Type 2 response to a problem as would have been involved in generating a Type 1 response to it. Quite the opposite. Type 2 thinking may bring a different, wider range of neural resources to bear on the problem, and it involves engaging executive and working memory systems as well. The claim is merely that there are no subsystems designed *solely* to support Type 2 thinking.

On this view, then, 'System 2' is not a neural system but a new level of organization, formed by culturally transmitted habits which restructure the activities of the biological brain. In Dennett's phrase, it is a softwired 'virtual machine', like a computer operating system, running on the hardware of the biological brain (Dennett 1991, Ch. 7). If the old mind is a biological mind, then the new mind is a virtual one.

The reader may suspect some sleight of hand here. How could perceptual and imagistic feedback so radically enhance the problem-solving powers of the brain? After all, the knowledge that we draw on in Type 2 thinking is encoded in Type 1 memory systems and available to Type 1 thinking. Why can't Type 1 processes take care of everything? There are several points to make here. First, as Dennett observes, feedback may enable the integration of information from different mental subsystems. Subsystems that lack internal channels of communication can share information by generating speech or sensory imagery expressing it, thereby making it available to perceptual systems and, through them, to the rest of the mind (Dennett 1991). Natural language is ideally suited to this role of content integrator, since most mental subsystems have access to the language system (Carruthers 2006). Second, imagistic feedback is not random but intentionally controlled, directed to solving some specific problem and guided by learned procedures and tricks, as discussed earlier. We learn ways of constructing verbal arguments and exploiting sensory imagery, just as we learn to do maths, drive, or play tennis. Such learning, of course, involves myriad micro-changes to the biological brain, encoding the new beliefs, skills, and habits. Third, Type 2 processing enables us to exploit our existing knowledge in new ways. Our memories encode a vast amount of information, all potentially relevant to any problem we face. Autostimulation has a strong selectional effect. When we ask ourselves a question, many different items of knowledge compete for articulation in inner or outer speech. The ones that win then prime the

next round of selection, giving the edge to related items, and so on. Thus, by autostimulating we can hack a path through the informational jungle, making new connections and arriving at new conjectures. Of course, many paths turn out to be dead ends, but with persistence and self-criticism we can find good ones.

To sum up so far: There is robust evidence for a qualitative distinction between two types of thinking, intuitive and reflective. This distinction is best interpreted as one between autonomous processes and intentional reasoning. Autonomous processes can guide everyday behavior in familiar environments, but intentional reasoning is needed to deal with novel or complex problems. It involves creating overt representations, questioning ourselves, imagining relevant scenes and objects, and constructing arguments in inner speech. The objects and imagery produced act as autostimulations, providing fresh inputs to our autonomous processes, which may then generate a response in the form of more inner speech, other sensory imagery, or an emotional reaction. This reframes the problem or provides a partial solution to it, and in turn acts as a further autostimulation, and so on. In this way, by engaging in cycles of autostimulation and response, we can work our way through problems that would otherwise be beyond us. Culturally transmitted habits of autostimulation create a new level of mental activity, a virtual mind, which engages in reflective thinking. It is by installing this virtual system in our heads that we come to approximate to general intelligence.

3.6 Enhancing Human Intelligence

What implications does this dual-minds view have for the project of artificially enhancing human intelligence? The first thing to ask is which system we are thinking of enhancing: the biological mind or the virtual mind? The methods would need to be very different. Enhancing the biological mind would mean directly interfering with the hardware of the brain. We might seek to boost our cognitive functioning with nootropic drugs, neurostimulation, or genetic manipulation. We might extend our perceptual capacities by hooking up artificial sensors to our sensory cortices, relying on the brain's plasticity to extract the information they supply. More ambitiously, we might create artificial cognitive subsystems, which interface with our biological ones. These would probably have to be self-organizing systems, which could be implanted early in life and grow alongside the biological ones, forming complex low-level connections with them. None of these technologies will be easy to develop, and installing them will require detailed understanding of brain functioning and development.

Enhancing the virtual mind is a completely different matter. Indeed, the virtual mind is itself a cognitive enhancement – a set of tricks for extending the powers of the biological brain, often through the use of artefacts. These tricks created the new human mind, with its powers of hypothetical thinking and creative problem-solving, and it is very tempting to link their emergence with the 'cultural explosion' 30–60,000 years ago, when art, religion, and complex technology first appeared

(Mithen 1996). (We might say that the first technological singularity occurred in the Upper Palaeolithic.)

Moreover, the virtual mind itself can easily be enhanced. On a software level, we can learn new reasoning techniques – new procedures for constructing arguments, doing calculations, making decisions, and so on. Much of human education, formal and informal, is concerned with this kind of enhancement. Adding new hardware is easy too. Because we have internalized many tricks of autostimulation and added new private ones, we tend to think of our conscious minds as essentially private (the streams of consciousness in our heads) and to suppose that enhancing them would require tinkering with our brains. But this is to over-emphasize an incidental feature of intentional reasoning. From a functional perspective, the autostimulatory routines we run in our heads are on a par with public ones involving the manipulation of artefacts, such as writing or sketching. In both cases we intentionally produce and manipulate artefacts and symbols in order to transform complex problems into simpler ones that our biological minds can solve. Technology can vastly extend this process by transforming difficult abstract problems into easy practical ones. Think of using a calculator to solve a complicated mathematical problem. Instead of solving the maths problem itself, we now have to solve the much simpler problem of how to get the calculator to solve it. Again, the process is fundamentally autostimulatory. At each step the calculator provides us with new stimuli, creating new, simpler subproblems: which keys to press first, how to interpret the answer the calculator displays, what entries to key in next, and so on. The solutions to these simpler problems are provided by our Type 1 processes, and the solution to the whole problem is the product of cycles of internal Type 1 processing and external electronic processing, which constitute a temporally and spatially extended Type 2 process.

We also supplement our biological memories with external sources of knowledge, such as tables, reference books, and databases. Rather than posing a question to ourselves, we can consult an external resource, retrieving items of information for use in Type 2 reasoning. Again, from the perspective of the virtual mind there is no significant difference between biological memory and external information sources. Both are resources we intentionally access (by autostimulation in one case, with hands and eyes in the other), in the hope that they will yield reliable and relevant information. External resources merely expand the hardware on which the virtual mind is run.

These enhancements to the virtual mind are easy to install. The devices involved are designed to interface naturally with our biological minds through our hands and sense organs. We press the keys of the calculator and look at its display. So adoption is easy; we just plug in new cognitive aids via sensory interfaces. All that is required is some training in using the devices and interpreting their outputs. (We might be able to make the devices more efficient by developing interfaces that bypass the external organs, detecting motor commands in the brain and sending signals directly to afferent sensory pathways, but such shallow interventions would be relatively easy to accomplish.) For thousands of years, we humans have been enhancing our Type 2 thinking with artefacts, from writing instruments and abacuses through to

iPhones and smart glasses, and this sort of enhancement looks set to progress rapidly in coming decades.[9]

3.7 Artificial Intelligence

As I noted earlier, from a dual-system perspective, different traditions in AI can be seen as focusing on different mental systems: computational modelling of general intelligence focusing on System 2, and embodied, behavior-based approaches on System 1. The former project has proved notoriously intractable, and the present view of System 2 sheds some light on this. If System 2 is a virtual system, then in order to reproduce its powers, we would need to reproduce the powers of the biological mind, too – the vast suite of fast, automatic, intelligent subsystems that forms the engine of System 2 thinking. To adopt a top-down approach is to put the cart before the horse, like trying to create an operating system without having the hardware to run it on. While a virtual mind may be easy to enhance, it is difficult to create.

In principle, no doubt, general intelligence could be modelled directly from the top down, perhaps even in computational terms, but it would be a formidable challenge. (If this isn't obvious, consider that it would involve, among other things, finding ways of representing all the diverse kinds of Type 1 knowledge in a format that allows for their integration in reasoning; devising procedures for rapidly retrieving contextually relevant items from a vast knowledge base; and creating a powerful general reasoning system that can perform a wide range of operations, including belief fixation and updating, decision making, planning, causal reasoning, mentalizing, language processing, abductive inference, and creative thinking.) Moreover, it is unclear what the target of the project would be. It is tempting to take our Type 2 thought processes as the paradigm of general intelligence, but we should not idealize them. Human Type 2 thinking is shaped by many contingent factors: by the nature and capacities of the specialist subsystems that drive it, by the cultural resources available for its programming, and by individual differences in the way we conduct it. If we were trying to model general intelligence computationally, it is not obvious that we should focus on our own idiosyncratic, species-specific and culture-specific form of it (unless, of course, we want to create artificial versions of ourselves).

A more practicable approach to creating general intelligence would be to work from the bottom up, creating independent creatures with Type 1 minds and coaxing them into developing Type 2 minds for themselves. We would need to equip them

[9] For careful exploration of how internet technology is extending and transforming human agency and cognition, and the costs and benefits involved, see Clowes 2017, 2019. Clowes stresses that although current developments have novel features, they continue a long-established process through which the human mind has been re-shaped and enhanced through interactions with material culture.

with goals, social instincts, suites of perceptual, cognitive, and motivational systems, and a communication system. By tuning their goals in the right way, we might get them to start cognitively stimulating each other and then autostimulating, working their way gradually toward explicit Type 2 thought. It is unlikely, however, that we could ensure this outcome through engineering alone. Our creatures would need to develop social institutions and cultural practices in order to sustain and transmit the skills and knowledge required for Type 2 thinking. As deliberate designers we could only take the process so far, but as guides and teachers we could take it further, sharing the mental software that has made us who we are. We might train our creatures, as we train children, providing scaffolding that helps them learn how to think. 'What might help?' 'What do you need to know?' 'Could you look at it differently?' 'What if you did this?' Our interactions with AIs may be much like those with precocious children.[10]

It may be, then, that the best way to create general intelligence will be to create beings who can create it for themselves. If so, then AIs will also have two minds, though the shape of both will probably be quite different from ours. The form of Type 2 thinking is determined by the nature of the autostimulatory mechanisms employed (the language system, perceptual and imagistic abilities, working memory capacity, and so on), and the virtual minds of AIs might be much richer and more complex than ours.

3.8 The Risks of Enhancement and AI

Speculation about enhanced and artificial intelligence soon turns to concerns about the risks involved, and I shall close this chapter with some remarks on this from a dual-minds perspective.

A common worry is that, having embarked on the creation of artificial intelligence, we may lose control of the process. Our creations may take control of their own development, pursue their own projects, and become indifferent or hostile to us. I think this is alarmist. For AIs to take control in this way, they would need to be capable of flexible, creative thinking of the Type 2 kind. They would need to be able to set themselves new goals, evaluate hypothetical scenarios, plan ahead, and much more. But, as we have seen, such abilities won't be easy to engineer, and a more feasible strategy will be to create artificial creatures with animal-like intelligence, and then help them to bootstrap themselves into general intelligence though cultural processes. This is unlikely to be a fast or straightforward process. We worry about AIs developing rapidly and escaping our control, but it is more likely that we shall have to nurture them laboriously through a long childhood, both as an artificial species and as individuals. Before we have to deal with super-intelligent AI overlords,

[10] For a related perspective, which explores the role of language use in developing a variety of higher cognitive functions in robotic systems, see Mirolli and Parisi 2011.

we shall probably have to spend many years dealing with demanding, reckless, accident-prone, and occasionally brilliant artificial children.

There is, however, another way in which we may cede control to technology, which is a much more pressing concern. I stressed how easy it is to enhance our virtual minds, using artefacts to transform problems and to supplement our biological memories. We have been enhancing and extending our virtual minds in this way for thousands of years, and our modern minds are heavily dependent on external support. (Think what effect the loss of your phone would have on your ability to do your job or organize your life.) Modern technology is accelerating this process, however, offering increasingly powerful new cognitive aids. Programming our biological brains to support Type 2 thinking is a laborious business, which involves mastering complex reasoning procedures and memorizing vast amounts of information. Computer technology offers shortcuts. Instead of learning to do long division, we can learn to use a calculator; instead of memorizing historical facts, we can learn to access an online encyclopaedia; instead of memorizing spellings, we can learn to run a spellcheck program. Technology looks set to supply us with ever more powerful shortcuts like this, allowing us to offload cognitive drudgery onto electronics in the way that previous generations offloaded manual labor onto mechanical appliances.

We can also expect technology to give us many completely new capacities, supplementing our biological minds with external modules, tightly linked via sensory interfaces. We shall be able to query these modules for information, entertainment, and motivational stimuli, and use them to make visual, aural, and tactile contact with far-off people and places. We can expect our conscious minds to be radically enriched, allowing us to develop new ways of working, socializing, and loving.

The advantages of all this are obvious, and we shall probably find them impossible to resist. (Why should a lawyer spend years studying case law if they can buy a tiny earpiece that will instantly retrieve contextually relevant data as needed and feed it to them?) But the dangers are obvious too. Making our conscious minds dependent on external electronic hardware as well as our biological brains will be a risky business. Our brains are robust, well-protected organs, which are the product of millions of years of natural R&D and have a remarkable capacity for self-repair. Electronic devices are far more vulnerable. A solar flare might knock them out and leave us cognitively disabled. And if they fail, it won't be easy to fall back on older technology. (Who now knows how to use a slide rule?)

More worryingly perhaps, we shall be at the mercy of those who control the technology. Having offloaded so much of our skill and knowledge, we won't have the resources to assess the value of the information and guidance we are fed, and those who control the feed will be able to manipulate the rest of us. We are already seeing something like this in the use of social media bots to manipulate opinion during elections. Seemingly relevant images and bits of information pop up on social media, just as thoughts pop into our heads, and it is easy to let them guide one's thoughts and decisions. Imagine having a host of similar bots whispering in your ear, guiding your work, your social relations, your personal life, your very thinking.

The moral, then, is that it is not the master AIs we should worry about but the servant ones. We may end up developing our virtual minds to the point where they are no longer really ours, no longer tethered to our biological minds and to the purposes and values rooted there. This is the paradox of the virtual mind. In learning how to manipulate our biological minds and create virtual minds for ourselves, we risk undermining the locus of purpose and control that our biological minds sustained. It is the price of being creatures with two minds.[11]

References

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence, 47*(1–3), 139–159. https://doi.org/10.1016/0004-3702(91)90053-M.

Carruthers, P. (2006). *The architecture of the mind: Massive modularity and the flexibility of thought*. Oxford: Oxford University Press.

Carruthers, P. (2009). An architecture for dual reasoning. In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 109–127). Oxford: Oxford University Press.

Carruthers, P. (2015). *The centered mind: What the science of working memory shows us about the nature of human thought*. Oxford: Oxford University Press.

Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual-process theories in social psychology*. New York: Guilford Press.

Chen, S., & Chaiken, S. (1999). The heuristic-systematic model in its broader context. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 73–96). New York: Guilford Press.

Clowes, R. W. (2017). Extended memory. In S. Bernecker & K. Michaelian (Eds.), *The Routledge handbook of philosophy of memory* (pp. 243–254). Abingdon: Routledge.

Clowes, R. W. (2019). Immaterial engagement: Human agency and the cognitive ecology of the internet. *Phenomenology and the Cognitive Sciences, 18*(1), 259–279. https://doi.org/10.1007/s11097-018-9560-4.

De Neys, W. (Ed.). (2018). *Dual process theory 2.0*. New York: Routledge.

Dennett, D. C. (1991). *Consciousness explained*. New York: Little, Brown.

Dennett, D. C. (2017). *From bacteria to Bach and back: The evolution of minds*. New York: W.W. Norton.

Descartes, R. (1984). *The philosophical writings of Descartes: Volume 2* (Trans., J. Cottingham, R. Stoothoff, & D. Murdoch). Cambridge: Cambridge University Press.

Diaz, R. M., & Berk, L. E. (Eds.). (1992). *Private speech: From social interaction to self-regulation*. Hillsdale: Lawrence Erlbaum.

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *The American Psychologist, 49*(8), 709–724.

Evans, J. S. B. T. (1989). *Bias in human reasoning: Causes and consequences*. Hove: Lawrence Elrbaum Associates.

Evans, J. S. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review, 13*(3), 378–395.

[11] I am grateful to Wim De Neys and two anonymous reviewers for comments on earlier drafts of this chapter. Some of the ideas in the chapter were first sketched in an interview for *Interalia* magazine (https://www.interaliamag.org/interviews/keith-frankish/), from which a few sentences are drawn.

Evans, J. S. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement*. Hove: Psychology Press.
Evans, J. S. B. T. (2010). *Thinking twice: Two minds in one brain*. Oxford: Oxford University Press.
Evans, J. S. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove: Psychology Press.
Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241. https://doi.org/10.1177/1745691612460685.
Frankish, K. (1998). Natural language and virtual belief. In P. Carruthers & J. Boucher (Eds.), *Language and thought: Interdisciplinary themes* (pp. 248–269). Cambridge: Cambridge University Press.
Frankish, K. (2004). *Mind and supermind*. Cambridge: Cambridge University Press.
Frankish, K. (2009). Systems and levels: Dual-system theories and the personal-subpersonal distinction. In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 89–107). Oxford: Oxford University Press.
Frankish, K. (2010). Dual-process and dual-system theories of reasoning. *Philosophy Compass, 5*(10), 914–926. https://doi.org/10.1111/j.1747-9991.2010.00330.x.
Frankish, K. (2016). Illusionism as a theory of consciousness. *Journal of Consciousness Studies, 23*(11–12), 11–39.
Frankish, K. (2018). Inner speech and outer thought. In P. Langland-Hassan & A. Vicente (Eds.), *Inner speech: New voices* (pp. 221–243). Oxford: Oxford University Press.
Frankish, K., & Evans, J. S. B. T. (2009). The duality of mind: An historical perspective. In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 1–29). Oxford: Oxford University Press.
Gigerenzer, G. (2010). Personal reflections on theory and psychology. *Theory & Psychology, 20*(6), 733–743. https://doi.org/10.1177/0959354310378184.
Gleick, J. (1992). *Genius: The life and science of Richard Feynman*. New York: Pantheon Books.
Heyes, C. M. (2018). *Cognitive gadgets: The cultural evolution of thinking*. Cambridge, MA: Harvard University Press.
Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.
Keren, G., & Schul, Y. (2009). Two is not always better than one. *Perspectives on Psychological Science, 4*(6), 533–550. https://doi.org/10.1111/j.1745-6924.2009.01164.x.
Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review, 118*(1), 97–109. https://doi.org/10.1037/a0020762.
Melnikoff, D. E., & Bargh, J. A. (2018). The mythical number two. *Trends in Cognitive Sciences, 22*(4), 280–293. https://doi.org/10.1016/j.tics.2018.02.001.
Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*(02), 57–74. https://doi.org/10.1017/S0140525X10000968.
Mirolli, M., & Parisi, D. (2011). Towards a Vygotskyan cognitive robotics: The role of language as a cognitive tool. *New Ideas in Psychology, 29*(3), 298–311. https://doi.org/10.1016/j.newideapsych.2009.07.001.
Mithen, S. J. (1996). *The prehistory of the mind: A search for the origins of art, religion and science*. London: Thames & Hudson.
Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review, 11*(6), 988–1010.
Osman, M. (2018). Persistent maladies: The case of two-mind syndrome. *Trends in Cognitive Sciences, 22*(4), 276–277. https://doi.org/10.1016/j.tics.2018.02.005.
Pennycook, G., Neys, W. D., Evans, J. S. B. T., Stanovich, K. E., & Thompson, V. A. (2018). The mythical dual-process typology. *Trends in Cognitive Sciences, 22*(8), 667–668. https://doi.org/10.1016/j.tics.2018.04.008.
Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology, 19*, 123–205.
Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3–22.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah: Lawrence Erlbaum Associates.
Stanovich, K. E. (2004). *The robot's rebellion: Finding meaning in the age of Darwin*. Chicago: University of Chicago Press.
Stanovich, K. E. (2009a). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 55–88). Oxford: Oxford University Press.
Stanovich, K. E. (2009b). *What intelligence tests miss: The psychology of rational thought*. New Haven: Yale University Press.
Stanovich, K. E. (2011). *Rationality and the reflective mind*. New York: Oxford University Press.
Stanovich, K. E., & Toplak, M. E. (2012). Defining features versus incidental correlates of type 1 and type 2 processing. *Mind & Society, 11*(1), 3–13. https://doi.org/10.1007/s11299-011-0093-6.
Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*(05), 645–726.
Steels, L., & Brooks, R. A. (Eds.). (1995). *The artificial life route to artificial intelligence: Building embodied, situated agents*. Hillsdale: Lawrence Erlbaum Associates.
Vygotsky, L. (1986). *Thought and language* (Trans. & Ed., A. Kozulin). Cambridge, MA: MIT Press.
Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, MA: MIT Press.
Winsler, A., Fernyhough, C., & Montero, I. (Eds.). (2009). *Private speech, executive functioning, and the development of verbal self-regulation*. Cambridge: Cambridge University Press.

Keith Frankish is honorary reader in philosophy at the University of Sheffield, UK; visiting research fellow at The Open University, UK; and adjunct professor with the Brain and Mind Programme in Neurosciences at the University of Crete, Greece. He is the author of *Mind and Supermind* (2004) and *Consciousness* (2005), as well as numerous journal articles and book chapters. He is the editor of *Illusionism as a Theory of Consciousness* (2017) and co-editor of *In Two Minds: Dual Processes and Beyond* (with Jonathan Evans, 2009), *New Waves in Philosophy of Action* (with Jesús Aguilar and Andrei Buckareff, 2010), *The Cambridge Handbook of Cognitive Science* (with William Ramsey, 2012), and *The Cambridge Handbook of Artificial Intelligence* (with William Ramsey, 2014). His research interests include the nature of phenomenal consciousness, the psychology of belief, and dual-process theories of reasoning.

Chapter 4
Does Artificial Intelligence Have Agency?

Danielle Swanepoel

4.1 Introduction

Knowing the agency-status of individuals is what underpins the way an individual is treated in society, in the legal system, in the educations system, in religious institutions etc. For example, we are not likely to consider a two-year old child an agent and so therefore, we do not hold a child accountable for her actions, and furthermore we are obligated to treat the child a certain way – as someone of deserving moral consideration.

Insofar as an individual acts in the world and insofar as these actions impact others and the environment, determining the agency-status of that individual is what often decides how we reward or punish that individual for their behavior. We have reached the point in time where artificial intelligence systems are performing actions in the world, such that they impact others and the environment, thus a discussion must be had if these things have agency.

What is an agent then? In its simplest form, an agent is a thing which performs *intentional actions*. What constitutes an intentional action is, roughly, that the action is something an agent wishes or desires to do.

An unintentional action is one in which the agent has little to no control over. Think about the difference between intentionally moving your leg to walk and the unintentional movement of your leg twitching while you sleep. Essentially, *control* is an important feature of intentional action and is considered one of the hallmark features of agency. However, it is possible for someone to perform an action (seemingly under their control), but against their will. This would look like someone who is forced, at gunpoint, to relinquish their wallet to a mugger. They have control over

D. Swanepoel (✉)
SolBridge International School of Business, Daejeon, South Korea

Faculty of Humanities, University of Johannesburg, Johannesburg, South Africa
e-mail: Dswanepoel@solbridge.ac.kr

R. W. Clowes et al. (eds.), *The Mind-Technology Problem*, Studies in Brain and Mind 18, https://doi.org/10.1007/978-3-030-72644-7_4

their bodily movements enough so to retrieve the wallet from their pockets successfully, but the action itself is born from coercion and therefore their autonomy is undermined. *Autonomy* is considered a second important feature of intentional action and is considered one of the hallmarks of agency.

If an agent is that thing which performs intentional actions, then an agent is also that thing which has a complex range of desires, beliefs, and deliberations which are involved in bringing those intentional actions about. So, then, an account of agency would need to show how, or under what conditions, these complex range of desires, beliefs and deliberations occur. It is not enough to say that there must be control and autonomy for intentional action to occur. One can think of higher-order animals who seem to have control and autonomy over their actions, but it is debatable if we would consider their actions as those of agents.

If we have this type of fleshed-out account, we are then in the position of knowing what counts as an agent and what falls short of being considered an agent. Simply put, a fleshed-out account like this provides us with a diagnostics tool – a set of criteria that need to be met in order to determine agency-status.

Over the last decade or so, interest in artificial intelligence (AI) has piqued. The reasons are many, too many to mention here. Suffice to say, one of the main reasons is *prevalence.* AI pervades almost every aspect of our lives; from the devices we use to communicate with others to the way we open doors, or switch on lights. AI has made our lives a lot easier with regards to communication, connectivity, movement, manufacturing, global positioning etc. AI is not just making great strides on the mechanical side of things (hardware), but also with the *operations* behind these mechanical things (software). It can calculate, weigh up different options and choose the best one which will ensure the best possible outcome, right a wrong, troubleshoot, solve a problem etc. These machines perform actions, and in some instances, these actions seem intentional. But does it have agency?

I answer this by examining four different accounts of agency. First, I extract the main features of each account and develop what I call common-ground agency. Common-ground agency essentially consists of the features of agency which are shared by all four accounts (Sect. 4.1 and 4.2). Next, I examine the relationship between AI and these features of common-ground agency (Sect. 4.3). I argue that AI fails to satisfy the four features, and because of this, we should be careful to think of AI as having agency (Sect.4.4). Finally, I explore what this means for phenomenal consciousness (Sect. 4.5).

4.2 Four Accounts of Agency

According to the standard conception of agency, briefly, if something acts with intention, then it is an agent.[1] The 'intentional' bit about intentional action is what allegedly sets humans apart from lower-order animals. If we remove the 'intentional' from the equation, then action is merely behaviour (Mele and Moser 1994). Therefore, according to this standard conception of agency, anything which performs intentional actions, is an agent. Notably, an AI system performs actions, and it can appear as if to do so with intention, especially for those who are unfamiliar with its programming. But there should be a distinction between something that looks a certain way to that thing actually being that. For example, one can act as if one is in love without ever actually being in love.

What we do when we tie action and intentionality together in such a fashion is essentially begin to tell the story of agency. An agency is not just something which just acts. An agent is something that performs actions intentionally. Understanding agency in this way sets out a necessary dependence between agency and action. Agency is necessarily dependent on the capacity for intentional action. Without intentional action and the capacity thereof, one is not an agent.[2] In Mele and Moser's account, intentional actions are those which "are intentional in virtue of their being suitably guided by an intention-embedded plan" (Mele and Moser 1994, p. 46). An intention-embedded plan is one which identifies goals and the actions to get there – thus satisfying the capacity of being able to deliberate on actions as intentional.

But there must be more than just the mere coupling of these two things surely? There is a lot that underpins this coupling and is something that is explored in most accounts of agency or action. Something I will briefly outline below by looking particularly at accounts offered by Harry Frankfurt (1971, 1999), Michael Bratman (2001, 2007), Christine Korsgaard (2009), and David Velleman (2000, 2006). What I hope to achieve is to extract common ideas which are shared by all four accounts. I do this in order to identify a set of criteria for agency, which most could agree upon.

4.2.1 Agents Have Desires

Frankfurt's account of agency sees intentional action as being intimately linked to the conception of desires, such that different order desires need to align in order for an action to count as one of a full-blooded agent. He conceives of an agent as having first-order and second-order desires and defines these first- and second-order desires

[1] Proponents of this view include, but are not limited to, Davidson (2001), Anscombe (2000), Ginet (1990), Brand (1984), and Mele and Moser (1994).

[2] Of course, one could argue that even through inaction we are agents (in the sense that even if we do nothing, we are still abstaining, which is an action in and of itself) and thus, such a strong dependency relationship between agency and intentional action is debateable.

as follows: a first-order desire is a desire or wish to perform such and such action. A second-order desire is a desire to have or not to have some first-order desire (Frankfurt 1971, p. 7). Frankfurt (1971, p. 8) further discusses the concept of an agent's *will* being the desire by which the agent is motivated to take action – thus providing a motivational drive behind intentional action. There appears to always be some kind of tension between the first-order and second-order desires and Frankfurt considers the second-order desire to be the "effective desire" (Frankfurt 1971, p. 9) – such that the second-order desire should agree with the first-order desire in order for the desire to be truly autonomous and of the agent's will. For example, an alcoholic might seek to refrain from drinking alcohol. Her first-order desire is to have a drink and her second-order desire is to not want to have that first-order desire to have a drink. If she proceeds to have a drink, then arguably her will is not free, because her second-order desire did not have an effect on her first-order desire.

But there's more at play than just having the right kind of desires aligning, reasons for actions lie in what is important for us, and a person "identifies himself with what he cares about in the sense that he makes himself vulnerable to losses and susceptible to benefits depending upon whether what he cares about is diminished or enhanced" (Frankfurt 1988, p. 260). And so, if the Frankfurtian agent believes it important that he overcome his desire for alcohol, then by not doing so, diminishes his sense of agency. It is, after all, that "by caring about things that we infuse the world with importance" (Frankfurt 2004, p. 23).

Admittedly, this is a very brief outline of Frankfurt's account and more is needed to do it justice – but this minimal account should suffice for the purpose of this paper. The following core features thus emerge as fundamental in Frankfurt's account of agency:

(a) An agent (person) is an individual who desires to have the desires she has and confers value through what it is that she cares about.
(b) An agent (person) is an individual who is capable of knowing what it is she desires by being self-reflective.
(c) An agent (person) is an individual who has "freedom of the will" (Frankfurt 1971, p. 14).
(d) An agent is one who is capable of volitional acts.

4.2.2 Agents Plan Ahead

Frankfurt's agent is one which aligns desires and knows what it is she cares about which informs her actions. Missing from that account however is how the agent unpacks her desires or how the agent goes about understanding what it is she cares about. If an agent is that thing which performs intentional actions, then an agent is that thing which *deliberates* about performing the intentional actions in the future. In order to deliberate about performing intentional actions, an agent must engage in a process of planning to successfully execute these intentional actions. In the event

that no planning takes place, the action would then be considered more spontaneous than intentional. Even the most mundane of intentional actions takes some form of planning. Think about making a cup of tea. This process involves sourcing a tea-bag, milk, sugar, and a cup. It involves boiling water etc.; This simple process takes planning.

Bratman argues for something he calls planning agency. Planning agency is agency directed at goals, which involve beliefs, desires, intentions and attitudes about the future. Planning agents comprise of planning structures, planning states, and means-end coherence. Further, the means-end coherence and consistency build temporally extended agency, which is a primary feature in Bratman's account of agency. Bratman's agent is that individual who perceives herself as having a future in which she either wants to, or is expected to perform actions, further understanding that actions bring about outcomes. Therefore, to accomplish goals or a set of outcomes, an agent would need to plan the route to get there. This requires that the planning is able to *reflect* upon what it is she wants to achieve and the best way to get there. The following features are fundamental in Bratman's account of agency:

(a) An agent is one who plans and must be reflective too.
(b) An agent is one who is aware of her past and future.
(c) An agent is one who is guided by planning structures.

4.2.3 Agents Are Functionally Rational

So far, we have agents as those who have desires and who plan out actions in the future. But how do we make sure all this is done reasonably and rationally and should reasonableness and rationality even be a requirement? Possibly one of the most influential and widely referenced constitutive accounts of agency can be found in Christine Korsgaard's work: *Self-Constitution* (2009). Korsgaard's account is complex in that it establishes the construction of agency in action and the process of bringing about that action. Korsgaard argues that the function of action is to constitute agency – and by doing so, action constitutes the identity of a person. According to Korsgaard, to say that something belongs to a particular kind is to identify with the kind's teleological structure or organization. This would entail that belonging to a particular kind means to possess the right kind of function. It appears then that something is identified as a particular kind by what it does. For example,

> the function of a house is to serve a habitable shelter, and that its parts are walls, roof, chimney, insulation, and so on. Then the form of the house is that the arrangement of those parts that enables it to serve as a habitable shelter (Korsgaard 2009, p. 28).

The core notion here is that being a habitable shelter is constitutive of being a house. In the same vein, when Korsgaard suggests that action is self-constitutive, she is suggesting that through action, one is constructing agency and thereby telling us more about what action consists of, and what counts as action. If we feel compelled to ask why actions should constitute an agent, the simple answer is that they just

do – in the same way that being a habitable shelter is constitutive of a house. I am an agent by virtue of what I do.[3]

A second important aspect to understand of Korsgaard's account is the way the constitutive feature of agency is tied into the notion of practical rationality. The core notion here is that being a habitable shelter is constitutive of being a house and that "a thing's constitutive function or form *just is* its constitutive norm" (Silverstein 2016, p. 216). When asking why it is that we ought to have a roof on a house, the answer simply is that it is essential for the house to function as a habitable shelter. Therefore, those actions belonging to a good agent are those which adhere to, or are guided by, rationality.

The following features are fundamental in Korsgaard's account of agency:

(a) Agency is established through action
(b) A unified agent is developed through unified action guided by rationality norms (such as coherence and consistency).

4.2.4 Agents Have Drives

As things stand, our agent is one who has desires and plans and these elements are guided by rationality norms – this is what it means to be an agent. But if this is so, why do we care to make sure these desires and plans are guided by rationality norms? What is our motivation to be good agents? Velleman offers a constitutive account where a primary, non-reducible drive to make sense of what one is doing and to do only what makes sense is the hallmark of agency. Velleman presents this constitutive account in the following way: firstly, he highlights that what is wrong with the standard model of action is that it leaves the agent out of the story. He further identifies that there is a problem with a hierarchal model of action (such as proposed by Frankfurt, 1971, 1988, 1999), the problem being that as long as there is a favourable reaction to having one desire over another, that will inform the action. "It doesn't matter, in the hierarchal model, whether the subject is satisfied with his first-order motives because of depression or boredom or laziness, or, alternatively, because it is responding to their force of reasons" (Velleman 2000, p. 14). Again, the agent is missing even in the hierarchal model. Because of this, Velleman proposes a constitutive account of agency – in which he establishes a constitutive aim of action – similarly to the way there is constitutive aim of belief.[4]

[3] A notion which Korsgaard entertains (but which has met with criticism because she is unable to provide a satisfactory account of this) is that it is possible that one can constitute the self badly (through failure to adhere to norms for instance) in the way that a house without a roof does not constitute a very habitable shelter.

[4] For more information on constitutive aims of belief, see Railton (2003). Railton presents his own constitutive account by arguing that belief is a propositional attitude which "cannot represent itself as unresponsive to – unaccountable to – […] truth" (Railton 2003, p. 297). The constitutive nature of belief is such that propositional attitudes which are not aimed at truth – or at least which are not

For Velleman, the constitutive aim of action is to *know what it is which we are doing*. In other words, if we were to infuse a creature of our own creation, for instance, with agency, we would need to design the creature "to gravitate towards knowing what they're doing, and they will only do those things which they have made up their minds that they're going to do, and so they will act by choice" (Velleman 2000, p. 26). Essentially, the action which is most intelligible for the agent to take will be the one that she is typically driven to choose. Similarly, to Korsgaard, normativity plays a crucial role. If we think of the drive for sense-making as guided by norms, then by accepting the process of sense-making, we are accepting the restrictions enforced by normativity.

To reiterate, an agent is one who always strives to do what makes the most sense to do – and the better the explanation of what it is we are doing is the better reason to do it. When a person takes herself by surprise with an action – perhaps when she acts instinctively – it is because she lacks the self-knowledge (knowing what makes sense) necessary to explain why she performed that action in the first place.

This drive to make sense of what it is we are doing – "the drive for sense-making" (Mitova 2016) motivates us to act for reasons. Further, this drive for sense-making ensures authentic, autonomous action. Velleman argues that we can think of autonomy as "conscious control over one's behavior", that is being conscious of one's behavior and of being in control of one's behavior.

The following core features are fundamental in Velleman's account of agency:

(a) An agent is reducible to a functional mental state from which the agent cannot disassociate herself.
(b) An agent is an individual who seeks to make sense of herself and the world.
(c) An agent plays a causal, authoring role in action.
(d) An agent critically reflects and deliberates upon the decision process (Velleman 1992, p. 477)

These four accounts share some common features. I imagine all four theorists would agree that an agent should have the capability to reflect and deliberate. Reflection and deliberation seem necessary in being able to distinguish between first-order and second-order desires for example. They also seems important for things like planning actions as well as being guided by normativity. I imagine that if I am incapable of reflection and deliberation, my ability to be guided by rationality norms would be very much skewed.

A second feature which I take to be important in all accounts is the conception of time. Particularly with the agent who is grappling with overcoming addiction for example, or the agent who intends to perform an action in the future. It certainly is an important element for having the drive for sense-making where the agent is examining themselves as existing in the present and existing in the future. What

responsive to admitted evidence – are not beliefs. Further, that beliefs are partially constitutive of agency, such that if an individual does not have belief, they fail to be considered agents.

makes sense to do implies an immediate future where one can enact what it is that makes sense to do.

A third shared feature is an awareness of an environment. This agent is one who grapples with aligning desires, who plans a future based on past and present events and who adheres to norms to the best of her ability and this assumes, at a minimum, someone who can engage with their environment and to understand that they are distinct from others in their vicinity – distinct from other agents, non-agents, and objects.

Finally, a feature also shared by all is the notion of freedom of choice. With regards to desires, we are free insofar as those different order desires align. When it comes to planning, our plans fall flat and are not ours as agents when we have no free-will to choose what it is we wish to do. If we adhere to norms without free-will, then arguably we are not adhering to norms, we are just simulating norm-adherence. In the next section, I explain these ideas further and show how these ideas form the basis for what I call *common-ground agency*.

4.3 Common-Ground Agency

In the last section, I briefly outlined some common features shared in four prominent accounts of agency. Here, my hope is to establish four minimal criteria which an individual would need to meet in order to be considered an agent. With the common-ground agency criteria in place, we can evaluate if individuals or entities which look and act like agents, are indeed agents. Entities which may fall under this description could be certain AI systems, robots, higher-order mammals etc. I am particularly interested in whether AI systems can meet these criteria and whether we are in the position to grant them agency-status. I argue that we are not there yet.

4.3.1 Reflection and Deliberation

The account of common-ground agency tells us that there are at least four features that need to be met in order for an entity or individual to be considered an agent. I identify the first feature as having the capacity to reflect and deliberate. An individual should both possess and identify with desires, beliefs and preferences as those are largely what guide and determine her actions. Identifying with these desires, beliefs and preferences is, in a sense, to establish that those desires, beliefs and preferences are your own and hold value for you. In Frankfurtian fashion, we, as agents, care about objects, relationships and events in our lives and we care insofar as we identify ourselves with what it is we care about. Furthermore, reflection and deliberation allows us to step outside of our own desires etc., and recognize that other agents have desires and preferences which may be different from ours and this too informs our actions greatly.

This possession of, and identification with beliefs, desires and preferences, feeds into choice and action, and arguably serves as a foundation for proper autonomous control over actions. If one were not to possess, nor identify, with these beliefs, desires, and preferences, arguably one would not be in the position to plan for the future, one would not be guided by norms and one would not have the drive for sense-making.

Drawing on empirical data, Cassidy et al. (2005) found that 3-year-olds find it difficult to move beyond their own desires and preferences to recognize the desires and preferences of another. Consistently, when a young child is asked to pick a gift for their parents, they would pick a gift that aligns with their own desires – not, potentially, what the parent may desire. In this instance, identification with one's own desires conflict with what it is one cares about. Picking a gift that you value according to your own desires and not according to the preference of the parent who you clearly value more than the fluffy toy you have chosen is to show an inconsistency in desires and preferences. Of course, adults do this too. Sometimes we pick gifts for our partners because we wish to use that object or to have that object for ourselves. The key distinction, however, is that as adults we are aware that this is what we are doing and so it is an intentional action that entails reflection and deliberation.

4.3.2 Awareness of Self in Time

The second feature of common-ground agency is that an agent is one who has a clear grasp of time. She is an individual who understands herself as having been present in the past, being here in the present, and seeing herself in the future. It goes without saying that Bratman's account of agency is most prevalent here. Bratman's account of agency identifies agents as those which take part in cross-temporal organization. It is a fundamental feature of agency that an individual understands herself as having been in the past and potentially existing in the future. This ensures that an agent can learn from past actions and to undergo effective deliberation in planning for future actions.

If we are to imagine an individual that lacks the concept of time, we are to imagine an individual who feels no sense of urgency, who is unable to effectively learn from past experiences nor is unable to make plans for the future. Adhering to rationality norms too requires, at a minimum, that actions are performed sequentially in order for a desired goal to be achieved. Making a cup of tea requires that the water is boiled first, for example. Graduating from a program requires applying to be admitted and attending classes, and passing exams etc. All of these usually follow a sequential order. Without the concept of time within which we frame our actions, we are looking at a group of acts haphazardly thrown together with no sense of direction or purpose.

4.3.3 Critical Awareness of Environment

As important as the awareness of time is, without it being coupled with the awareness of an environment, we end up having two disjointed concepts, especially for an agent. The third feature of common-ground agency is that an agent is one who is aware of her environment. She must be able to see objects and events in the world as making up part of the bank of options available to her, and based on this she must be able to evaluate her options and understand causal relations in the world. She needs to understand her actions as causal and as having effects. For instance, if action is what constitutes the self, then an individual would need to have relevant and necessary knowledge of surroundings to know which actions will result in the best possible outcome. First, desires and preferences are often those things which are aimed essentially at objects outside of ourselves which enhance our lives. Second, for us to confer value (in the Frankfurtian sense), we need to be able to identify things which we care about – this obviously requires awareness of one's environment.

4.3.4 Freedom of Choice

The fourth feature of common-ground agency is that an agent must have freedom of choice. This means an agent must be in the position to act the way they wish and desire without hindrances, without coercion, and without other restrictions which are beyond her control. How do we measure this though? Simply, and this is something I argue elsewhere (Swanepoel 2020), we should ask the question of the agent: could they have done *otherwise*? A metric of genuine norm-adherence can be whether deliberate norm-violation is possible.

An individual who chooses to stop at a red light, even though they could cross if they so wished, looks different from an individual who is chained to a post and unable to cross at the red light if she so wished. In both instances, the norm of not crossing at a red light is being honored. But one is genuine and the other is not. An agent therefore, is an individual which is capable of purposefully violating norms – for instance, norms set out by rationality. According to Frankfurt, a person is one who identifies with her volitions. If her volition is to purposefully violate a norm, then she is still in possession of her will and chooses to violate the norms freely. Norm-violation and the ability to violate norms should not be underestimated when we are seeking to establish an account of agency. The idea is this: one cannot be said to adhere to norms without equally having the capacity to violate those norms. If an individual is unable to violate a norm, then we cannot say, for certain, that norm adherence is taking place.

Why is this a necessary feature of agency? Arguably, an agent is one who adheres to norms (Korsgaard 2009). Agency is compromised if norms are violated – some argue that agency is diminished if norms are violated. If an account is only able to

explain norm-adherence – worse still – if it is only able to explain agency in terms of norm-adherence, then it cannot account for agency.[5] This is clearly problematic as there are many instances, in everyday life, where agents violate norms without undermining agency. Given this, it is important that a successful account is able to explain the possibility of norm-violation so that we can be sure that norm-adherence is present. If an account is unable to explain norm-violation, then it cannot be said to explain norm-adherence. This is directly related to autonomous control over one's actions. If we are not able to violate norms, then it does not follow that we can adhere to norms.

4.3.5 What this Means

I suggest that if an individual fails to have one of these four features of common-ground agency, she would fail to act in a way that is characteristic of an agent. Think, for a moment of a simple action you might perform in the next few hours – that of enjoying a meal. When you think about this you are juggling a few ideas, such as where this meal will take place (environment), what it is you would *like* to eat (self-reflective), when it is that this meal will take place (time) and finally, thinking of having dessert – which you promised yourself you wouldn't do (norm-violation).

Imagine three possible worlds where this is likely to take place. The first world is one where you knowingly don't exist in the next hour (an end of the world scenario). The second possible world is one where your actions don't count for you. The third possible world is one most similar to yours now where you will enjoy your meal and where your actions count as yours.

In the first possible world, knowing that you will not be around to enjoy a meal in the future will deeply impact your decision-making processes about the future. It would be irrational to plan a meal which is contradictory to the belief that you will not exist to enjoy that meal. In order for us to effectively make plans for the future, we need to have reasonable knowledge that we will exist in the future to enjoy the benefits of our planning. Otherwise, the planning is hypothetical and does not pertain to me. And if it doesn't pertain to me, why should I care? Note however that it is still possible, in the face of non-existence in the next hour, to still take comfort in imagining oneself with your family on spring vacation (served by your memories). It's not irrational to remember the past. But there is a difference between remembering a past event and planning a future when there isn't one.[6]

The second possible world is different from the first because in this world you exist and continue to do so for the foreseeable future, but your actions are not yours

[5] For more on this, see David Enoch's *Agency, Shmagency: Why Normativity Won't Come from What Is Constitutive of Action* (2006).

[6] Many thanks to an anonymous referee for pointing at this in feedback.

and you feel no ownership of them. What this means is that benefits, and perhaps even negative outcomes that result from your actions are never experienced by you. It's almost like playing a slot machine at a casino and there's a payout every time, but you never receive it. In this world, your actions (if they can be called that) become meaningless and unmotivated. It wouldn't matter to you if your actions harmed others or helped others – why would they? They don't belong to you and they do not impact you.

In which world is adhering to rationality norms most likely to take place? Not in the first. Without a future, planning a future meal is incoherent. There is a deep inconsistency here between correct beliefs and your incorrect actions and epistemically this is problematic. But other than the epistemic problem at play here, planning itself is deeply impacted. Planning is only relevant if there is a future to speak of – otherwise the planning is just hypothetical nonsense. Decision-making and action also lack consistency and coherence in the second world. Why would I make decisions that pertain to me when the very actions I perform do not count for me? There is something deeply irrational about playing a slot machine that you know is paying out but you never receive any of the winnings (unless a world exists where you know the winnings are going to someone or something you care about). It proves to be a futile and useless act.

The only world where agency seems intact and where actions are guided by rational norms is the third possible world – the one you are in now. It is because this world is one where all four features of common-ground agency are accommodated.

4.4 An Analysis: AI and Common-Ground Agency

What I have available to me now is a minimum set of criteria that need to be met for an individual to be considered an agent. Now, I turn my attention to answering the question this paper has set out to establish: does artificial intelligence have agency? At this point, I must put forth two caveats. First, in this paper, I consider, primarily, more traditional rules-and-representations kinds of AI and will briefly consider more dynamic/ neural kinds of AI throughout.[7] The reason for this is, as I mentioned before, *prevalence.* The status quo is that most of the AI systems which we use daily still function in the realm of rules-and-representations and pattern-matching and I write this analysis of the possible agency-status of AI from this perspective.[8]

Second, the account of common-ground agency I propose above is comprised of shared features which I extracted from accounts aimed primarily at describing human agency and so is very human-centric. Because of this, even in the best-case

[7] Thank you to the editors for pointing out that I need to cater for more dynamic versions of AI.

[8] Of course, it is important to consider a future in which more substantially progressive AI may become the status quo and what I set out to achieve here today may not be steadfastly relevant in 50 years to come. But this is the case with almost all types of research which involves a fast-paced industry, such as technology.

scenario, it is highly likely that AI will fail to satisfy the set of criteria I have set out in virtue of their not being *human* and this could explain how and why they miss the mark. But, the situation is direr than this. I argue that even if we adapt the common-ground agency and convert it into something which is AI-friendly, AI would still fail to satisfy the criteria set out in common-ground agency. Further, that even if we cheat in favor of AI, we still don't get to a place where AI can be said to have agency- if anything, this provides compelling support that AI just can't get that grip on agency – at least not in the way that human beings can.[9] Here, I propose an AI-friendly version of the common-ground agency and analyze the agency-status of AI based on this.

In the account of common-ground agency presented in the above section, you will find words such as *awareness, sense, self, beliefs, desires,* and *understand.* Arguably, these human attributes and characteristics are what seem to push the human-agenda behind the account of common-ground agency. In order to convert the original account of common-ground agency into one that is more AI-friendly, I propose to revisit the terminology and concepts I use in the original account of common-ground agency, aiming to provide an account where the terminology and concepts are aimed at AI. I do this next.

4.4.1 Reflection and Deliberation – AI-Friendly

The first feature of common-ground agency is about an individual who possesses (and identifies with) *desires*, *beliefs* and *preferences*. Further, that she should have intimate knowledge of these things through identifying with them. *She* needs to further *understand* that these things play a part in determining *her* actions. The italicized words are words I associate with human attributes and characteristics. Arguably, we would be hard-pressed to think of AI as having desires, beliefs and preferences in the same way that human beings do. An AI-friendly version of this first feature may look something like this:

An individual satisfies this first feature if:

– it possesses necessary *goals, belief-networks* and preferences;
– it is able to compute that certain goals take precedence over others;
– *can compute* that these things play a part in determining *its* actions.

In AI programming, preference relations are about relations over outcomes (Pigozzi et al. 2016). An AI system is either programmed to have weak preference relations where one outcome is worth less than another, and an indifference preference relation, where one outcome is preferred equally with another, and finally, a strict preference relation, where one outcome has a far higher utility than another (Poole and Mackworth 2010). Human agents may not see preferences as strict value relations

[9] Thanks to an anonymous reviewer who noted that this move could be perceived as a cheat.

in numerical or programmable terms, as we have developed a set of heuristics over millennia to account for this, but, arguably what we value reflects as preferences and informs our actions. The same can be said about AI. The action which holds the highest value – perhaps the action which will maximize expected utility - is the action that will most likely take place if programmed to do so.

Georgeff et al. (1998) famously developed the so-called belief-desire-intention model (BDI). For Georgeff et al. (1998, p. 3), "In AI terms, Beliefs represent knowledge of the world. However, in computational terms, Beliefs are just some way of representing the state of the world". These Beliefs can either be represented as values or variables. Desires are viewed more commonly as goals. Again, a desire may also just be a value of a variable or some kind of record structure (Georgeff et al. 1998). If we had to think of desires as being present in AI, we should think of these desires as representing some or other end-state. Arguably, BDI models of AI do, to a degree, possess some semblance of beliefs and desires.

According to Guliuzza (2014, p. n.p) "Watson[10] was designed to change itself over time". For Watson to change itself, it must have some information available to it that informs it when it needs an upgrade or needs a modification. It may have access to data which indicates it is a system independent from another perhaps.

Takeno et al. (2005) developed a robot which is able to discriminate a mirror-image of itself from another identical robot. Because of this, it is likely that it does have some kind of recognition that it is different from another robot and further recognizes that it is an entity separate from another. Even though AI may have these capabilities of mirror self-recognition, it is unlikely that this equates to self-awareness in the same way human beings experience self-knowledge or sense of self.[11] Recognize though, that we do not require it to have the same kind of self-awareness as a human does – but it should have some semblance of goal-directed behavior. And so, with the most liberal interpretation of common-ground agency, and with a stretch of the imagination, AI can satisfy this feature of the AI-friendly version of common-ground agency.

4.4.2 Awareness of Self in Time – AI-Friendly

The AI-friendly version of the second feature holds that individual (or in this case, a system) computes its presence in a past time, its presence in the present moment, and its presence at a future time. Thus, time should be relevant as it pertains to its presence or existence. Minimally, to satisfy this feature, AI would need to have data available to it that conveys its presence over a range or period of time. We know that AI satisfies this feature with regards to the past because it has memory and capacity

[10] Watson is an AI system designed by IBM and is said to "understand all forms of data, interact naturally with people, and learn and reason, at scale" (IBM 2017).

[11] For more on this, see Alain Morin's "Self-recognition, Theory-of-Mind, and self-awareness in primates and right hemispheres" (2011).

to learn. Additionally, through if/ then statements of actions already performed, it can calculate the best possible outcome for future actions according to the variables available to it. However, its plans – if it can be said to have such things – are severely limited to what is has been programmed to do. Even if you consider a more advanced dynamic neural networks system which can be trained to learn sequential patterns or patterns which are time-varying or time-dependent and are perceived to have a more realistic performance, it struggles to do much more than make predictions about future actions it might take. This shows the absolute minimal engagement with this particular feature: predictions made from data gathered from a previous sequential pattern. If this feature were truly satisfied, AI would not just make predictions based on previous actions, but would somehow be invested, or identify with the future action. A future should *matter* for an agent and it is highly doubtful that the future matters to AI, even in the most simplistic way.

The reason the future should matter for an agent is related to the very notion of intentional action. It's difficult to provide an account of intentional action without appealing to a motivational or justificatory explanation as to why an agent would be compelled to perform the intentional action in the first place. A large part of this story of a motivational drive behind intentional action is dependent on the self being present in the future. It is difficult to imagine, let alone prove that AI systems (in the traditional sense or in the dynamic, progressive sense) possess this kind of identification with a future existence or presence. With no satisfactory account of motivation behind an AI's actions, this second feature of agency is barely met.

4.4.3 Critical Awareness of Environment -AI Friendly

The AI-friendly version of the third feature is the ability to compute inputs from the environment and, as a result, act in accordance with these inputs.[12] To have this feature, a system must be able to perceive objects and events in the world as making up part of the bank of options available to it. Based on this, it must be able to evaluate its options and compute causal relations in the world. It further needs to compute that its actions are causal and have effects.

Currently, in AI research, a lot of work is being done on so-called capture generation (Xu et al. 2015). Capture generation is the ability of an AI system to identify objects in its vicinity and to also "capture and express their relationships in natural language" (Xu et al. 2015, p. 1). An AI system such as Watson is able to analyze information provided to it in natural language and is able to establish relations between the keywords in the natural sentence. Thus, if it is able to convert the visual

[12] Thank you to an anonymous reviewer for pointing out that this feature, interpreted in this way, is setting the bar very low. It is possible that AI systems have "vision" but this vision is vastly different from human vision. And I agree with this. And this supports my argument further, that even if we proposed a completely adulterated version of the original common-ground agency, AI still fails to achieve agency-status.

representation into natural language, then it can engage with natural language, perhaps even in the same way as when it is asked a question in a natural language. A more relevant example of this can be found in self-driving cars. A self-driving car is designed and programmed to take in its environment and to react accordingly. It does this because it is able to analyze input from the environment.

Perception in AI systems takes place through the use of sensors which measure different aspects of the environment in a form that can be used by the AI system (Russell and Norvig 2010). There are two components to a sensory model: an object model (which describes objects in the world), and a rendering model (which describes the mathematical processes that result in the input from the environment) (Russell and Norvig 2010). Thus, it collects data of objects outside of itself. Further, AI systems are able to analyze the effects of its actions within this environment. Let's look at a simple example of the robotic vacuum cleaner. A robotic vacuum cleaner moves across a room in a certain pattern until it has effectively created a map of the room. It further can detect dirt and clean as needed. Once it is done with its chores, it returns to its docking station. It only returns to the docking station once it has vacuumed the dirt. To do this, it must have data that it has cleaned already and should not make a round of the room again (Russell and Norvig 2010, p. 39). Thus, according to the most generous interpretation of this feature of common-ground agency, AI systems can, in the best-case scenario, satisfy this feature.

4.4.4 Freedom of Choice – AI Friendly

The AI-friendly version of the last feature is that a system must be able to deliberately break the rules. True autonomy and control is reflected in the ability or capacity an agent has of being able to do that which is contrary to the acceptable, or contrary to the norm. If we have no choice and are forced to adhere to norms, then arguably our autonomy and control is compromised, if not absent entirely. Thus, an essential element of agency is the capacity for purposeful norm-violation as this is indicative of genuine norm-adherence. An illuminating way to know if purposeful and true norm-adherence is taking place is to establish that the agent has both the capacity to purposefully violate or adhere to the norm. Korsgaard explains the problem behind not being able to "if we cannot violate [a principle], then it cannot guide us, and that means that it is not a normative principle" (Korsgaard 1997, p. 321).

If norms are rules or principles which govern actions and decisions, then it means these rules and principles are showing you which actions and decisions are good, or better, or beneficial, compared to other actions and decisions which are bad, worse, or disadvantageous. It is not the case that norms just *are* good, better or beneficial actions and decisions. They are rules or principles which *govern* and *guide* a decision-maker through varying levels of acceptable and unacceptable actions and decisions. For an individual to be aware of both, she must know she is guided to perform an acceptable action knowing that an unacceptable one exists. In addition, that violating the norm is possible if the agent so desired, simultaneously realizing

that there are consequences for her decisions. Arguably, without the ability to purposefully violate a norm, it cannot be said that an individual performed the action autonomously. And so, for instance, something like malevolence is an act of norm-violation and is possible in almost all agents. But, as Pinker points out (1999, p. 16), "we are beginning to appreciate that malevolence […] does not come free with computation but has to be programmed in". It is questionable whether an entity which is *programmed* to accept norms is, in actual fact, capable of adhering to norms in the way described here.[13] This applies to both norm-adherence and norm-violation.

According to Castelfranchi et al. (1999, p. 364) "if the conventions and norms are hard-wired into the agent's protocols it cannot decide to violate the norms" and the reason it cannot make the decision to violate norms is because it doesn't understand the concept of norms as guiding principles for behaviour (Swanepoel 2020).

An AI system behaves according to a set of rules with which it was programmed. It becomes impossible then to derive norms from such facts embedded in the programming language; arguably, it is able, however, to derive further rules from the original rules. However, any behavior which simulates norm-accepting or norm-obeying is not *actual* norm-adherence. As things stand with rules-and-representations AI, AI does not undergo a deliberative, autonomous process (necessary for norm-adherence or norm-violation), because it cannot act against its programming (Swanepoel 2020). Even if we consider a more dynamic version of AI, it is difficult to show how this will result in purposeful norm-adherence. Here, AI fails the test.

4.5 Interesting Consequences

According to the features of common-ground agency, AI in its more traditional rules-and-representations version, and possibly, AI in its more dynamic version, fail to exhibit the features of common-ground agency. The result of this is that we should not (in the status quo) award AI systems agency-status, especially in the way required to perform autonomous, purposeful intentional actions. This claim has interesting ramifications for many discussions, but one I'll briefly focus on here is the ramifications about the complexities of phenomenal consciousness.

Intuitively, there exists a relationship between phenomenal consciousness and the four features of common-ground agency – one of correlation perhaps. The purpose of this paper was to answer whether AI systems have agency and I have concluded, given the failure of meeting all four criteria, AI systems do not. Here, I welcome the reader to consider a *possible* reason why they are unable to meet these criteria – that perhaps what is missing from AI systems is phenomenal

[13] Thanks for an anonymous reviewer who noted the relevance of Hume here. Hume's law (1962) encapsulates the problem nicely where this law states that norms cannot be derived from facts and that what we ought to do cannot be derived from what there is.

consciousness. Here I do two things, briefly: i) show that it is possible that phenomenal consciousness may underpin the features of common-ground agency, and ii) show that AI systems do not have phenomenal consciousness. Thereby concluding that perhaps this is the reason AI systems don't meet the requirements of common-ground agency, and iii) a surprising consequence of this finding could be telling of the nature of phenomenal consciousness.

Let's begin with (i). What is phenomenal consciousness? According to Coleman (2009: 83), when one learns to ride a bicycle,

> one tries to ride, and falls off. One tries again and falls again. Quickly one learns that a certain feeling- in fact that of toppling- means being imminently in for a tumble and a cut arm. Conversely, keeping the bike, however fleetingly, within the acceptable parameters of uprightness comes with its own distinctive new sensation: that of undisturbed equilibrium, of not being impinged upon by overpowering forces to either side (Coleman 2009, p. 84).

Rather too simply perhaps, phenomenal consciousness can be understood as the feeling of experience and it is something that still eludes philosophers and scientists alike. But Coleman's example is an excellent example of an instance of phenomenal consciousness. Human beings don't calculate the angle of the bicycle before it falls, nor we do work out the ratio of our weight versus leaning that might contribute to the bike falling over. We *feel* it as we ride. There's a very particular feeling to toppling over versus the feeling of "undisturbed equilibrium" and this is something we master through this *feeling* we have of ourselves riding our bikes.

What is the relationship between phenomenal consciousness however and common-ground agency? If we look again at the core features of common-ground agency, we will notice that, in order to satisfy some (if not all) of these criteria, arguably, phenomenal consciousness may be a requirement.

One of the underlying notions behind almost all of the features of common-ground agency is the ability to *identify with* things or to have a *sense*, such as identifying with desires and beliefs, identifying with the idea that one will exist in the future and identifying oneself as an actor in an environment where actions have consequences. Further, the notion of a *sense* of self, a *sense* of the environment and a *sense* of time all point to a subjective feeling, perhaps even going so far as to label it as qualia or the *feeling of what it is like* (Nagel 1974).

Typically, in studies about consciousness, there are two types or variants of consciousness: access consciousness and phenomenal consciousness. Access consciousness is the type of consciousness required to access information which allows us to make decisions, perceive the world and interact with the environment. Phenomenal consciousness is the experience and the 'raw feels' accompanying these interactions (Block 1995). It has also been described as consisting of the qualitative properties of experience, or qualia.

Let's move onto (ii). In the best-case scenario, truly advanced AI might possess access consciousness. We can look to examples of self-driving vehicles, autonomous drones, advanced robotics for examples of advanced technological systems having access to environments and the "self" therein.

But it is debatable that they will ever possess phenomenal consciousness which, interestingly, might be what it takes to satisfy the features of common-ground agency. I can't imagine Atlas (from Boston Dynamics) doing a cartwheel because it can *feel* when it is off-balance. Interestingly however, it is thought that "digital computers are more similar to minds than anything else known to us" (Piccinini 2007, p. 24) and for those searching for a mechanistic explanation of mind – including phenomenal consciousness – computational theories of mind (CTM)[14] are often explored, as: "CTM has become the mainstream explanatory framework in psychology, neuroscience, and naturalistically inclined philosophy of mind" (Piccinini 2007, p. 24).

Finally, let's talk about (iii) then. Even if CTM is the mainstream explanatory framework, it is not without its challenges. The ability to process notions such as one's existence in the future and to identify oneself as an actor in an environment where one's acts have causal consequences may require a little more than what is explicable by CTM. According to Tiziana Zalla (2007), representations of beliefs, desires and intentions (or, applicably, of time, environment, self, and norms) are accompanied by episodes of phenomenal consciousness which enjoy and serve a cognitive role. Further, "according to this hypothesis, p-consciousness was selected as a way of labelling some kinds of representations and to carry information about their origin" (Zalla 2007, p. 197) and that "accounts in terms of neural states that consider qualitative properties as epiphenomenally supervenient[15] can hardly explain the fact that they correspond to intentional features of the external environment" (Zalla 2007, p. 197).

CTM cannot explain why an individual might feel motivated or driven to perform one action over another, nor can it explain why an individual might plan the way they do or why an agent might feel compelled to violate a norm. CTM cannot tell me why I don't fall off my bike even though I'm actively not doing any calculations that might inform my actions. CTM cannot explain the feeling of experience. Therefore, as things stand, it seems unlikely that computationalism can provide a satisfactory account of phenomenal consciousness.

This leaves me with two possible outcomes: First, the features of common-ground agency do *not* require (in any sense) episodes of phenomenal consciousness and so, this cannot be the reason AI fail as agents. This option doesn't seem plausible: not only does it seem intuitively the case that phenomenal consciousness

[14] "Computational models of problem solving, where the problems to be solved are of the complexity of those solved by human beings" (McDermott 2007, p. 117) are most widely used by AI researchers and that "most AI researchers are computationalists to some extent" (McDermott 2007, p. 117).

[15] Epiphenomenalism is the view that mental events are somehow caused by physical events in the brain and that these mental events have no effect upon the physical. According to McDermott (2007, p. 130), for the computationalist, "phenomenal consciousness is the property a computation system X has if X models itself as experiencing things". Further, consciousness for the computationalist is the based on the notion that phenomenal properties somehow supervene on computational ones, so much so that being just the right kind of computation is enough to be considered conscious (Klein 2016).

plays an important part in what it means to be an agent, but there is also empirical support for the argument that phenomenal consciousness may be required (Damasio 2012; Zalla 2007; Tversky and Hard 2009; Thompson 2008).

Second, the features of common-ground agency do require episodes of phenomenal consciousness. This would then mean that *any* theory which can satisfactorily explain phenomenal consciousness and still fail to show that AI possess phenomenal consciousness can possibly provide the answer as to why AI fails to satisfy the features of common-ground agency. Hopefully, in the future, such a satisfactory account of phenomenal consciousness will be made available to us.

4.6 Conclusion

AI systems do act in the world. Because of this, determining their agency-status is one important contribution this paper makes. In this paper, I briefly unpacked several prominent theories in philosophy of action and agency and used core features of these accounts to construct the theory of common-ground agency. Within the constraints set out by common-ground agency, I asked the essential question of whether *Artificial Intelligence has agency*? I conclude, that given the status quo, they do not.

References

Anscombe, E. (2000). *Intention*. Cambridge: Harvard University Press.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18*, 227–287.

Brand, M. (1984). *Intending and acting*. Cambridge: MIT Press.

Bratman, M. (2001). Two problems about human agency. *Proceedings of the Aristotelian Society, 101*, 309–326.

Bratman, M. (2007). *Structures of agency*. New York: Oxford University Press.

Cassidy, K. W., Cosetti, M., Jones, R., Kelton, E., Rafal, V. M., Richman, L., & Stanhaus, H. (2005). Preschool children understanding of conflicting desires. *Journal of Cognition and Development., 6*, 427–454.

Castelfranchi, C., Dignum, F., Jonker, C., & Treur, J. (1999). Deliberative normative agents: Principles and architecture. In N. Jennings & Y. Lesperance (Eds.), *Intelligent agents VI, agent theories, architectures, and languages, Orlando: ATAL* (pp. 364–378). Berlin: Springer.

Coleman, S. (2009). Why the ability hypothesis is best forgotten. *The Journal of Consciousness Studies, 16*(2–3), 74–97.

Damasio, A. (2012). *Self comes to mind: Constructing the conscious brain*. New York: Random House.

Davidson, D. (2001). *Essay on actions and events*. Oxford: Clarendon Press.

Enoch, D. (2006). Agency, Shmagency: Why normativity won't come from what is constitutive of action. *Philosophical Review, 115*(2), 31–60.

Frankfurt, H. (1971). Freedom of will and concept of a person. *The Journal of Philosophy, 68*(1), 5–20.

Frankfurt, H. (1988). *The importance of what we care about*. Cambridge: Cambridge University Press.
Frankfurt, H. (1999). *Volition, necessity, and love*. Cambridge: Cambridge University Press.
Frankfurt, H. (2004). *The reasons of love*. Princeton: Princeton University Press.
Georgeff, M., et al. (1998). The belief-desire-intention model of agency. In J. Muller (Ed.), *ATAL* (pp. 1–10). Berlin: Springer.
Ginet, C. (1990). *On action*. Cambridge: Cambridge University Press.
Guliuzza, R., 2014. *IBM's Watson, designed to learn like a human.* [Online] Available at: http://www.icr.org/article/ibms-watson-designed-learn-like-human/. Accessed 21 July 2017.
Hume, D. (1962). *A treatise of human nature* (2nd ed.). London: Dent and Sons.
IBM. (2017). *IBM Watson.* [Online] Available at: https://www.ibm.com/watson/. Accessed 21 July 2017.
Klein, C. (2016). Computation, consciousness, and "computation and consciousness". In: M. Sprevak & M. Columbo (Eds.), *Routledge handbook of the computational mind* (p. n.p.). S.l.: S.n.
Korsgaard, C. (1997). The normativity of Instrumental reason. In G. Cullity & B. Gaut (Eds.), *Ethics and practical reason* (pp. 215–254). Oxford: Clarendon.
Korsgaard, C. (2009). *Self-constitution*. New York: Oxford University Press.
McDermott, D. (2007). Artificial intelligence and consciousness. In P. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness* (pp. 117–150). Cambridge: Cambridge University Press.
Mele, A., & Moser, P. (1994). Intentional action. *Nous, 28*(1), 39–68.
Mitova, V. (2016). What do I care about epistemic norms? In M. Grajner & P. Schmechtig (Eds.), *Epistemic reasons, norms and goals* (pp. 199–223). Berlin: De Gruyter.
Morin, A. (2011). *Self-recognition, theory-of-mind, and self-awareness in primates and right hemispheres.* [Online] Available at: https://www.researchgate.net/publication/242232773_Self-recognition_Theory-of-Mind_and_self-awareness_in_primates_and_right_hemispheres1. Accessed 1 10 2018.
Nagel, T. (1974). What it is like to be a bat. *The Philosophical Review, 83*, 435–450.
Piccinini, G. (2007). Computational explanation and mechanistic explanation of mind. In M. Marraffa, M. De Caro, & F. Ferretti (Eds.), *Cartographies of the mind: Philosophy and psychology in intersection* (pp. 23–36). Dordrecht: Springer.
Pigozzi, G., Tsoukias, A., & Viappiani, P. (2016). Preferences in artificial intelligence. *Annals of Mathematics and Artificial Intelligence, 77*(3–4), 361–401.
Pinker, S. (1999). *How the mind works*. London: Penguin Publishers.
Poole, D., & Mackworth, A. (2010). *Artificial intelligence: Foundations of computational agents*. London: Cambridge University Press.
Railton, P. (2003). On the hypothetical and non-hypothetical in reasoning about belief and action. In *Ethics and practical reason* (pp. 53–80). Oxford: Clarendon Press.
Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach*. Hoboken: Pearson Education.
Silverstein, M. (2016). Teleology and normativity. In R.-S. Landau (Ed.), *Oxford studies in meta-ethics* (pp. 214–240). Oxford: Oxford University Press.
Swanepoel, D. (2020). The possibility of deliberate norm-adherence in AI. *Ethics and Information Technology.* https://doi.org/10.1007/s10676-020-09535-1.
Takeno, J., Inaba, K., & Suzuki, T., (2005). *Experiments and examination of mirror image cognition using a small robot*. In The proceedings of the 6th IEEE international symposium on computational intelligence in robotics and automation, Volume CIRA 2005, pp. 493–498.
Thompson, E. (2008). Representationalism and the phenomenology of mental imagery. *Synthese, 160*(3), 397–415.
Tversky, B., & Hard, B. (2009). Embodied and disembodied cognition: Spatial perspective-taking. *Cognition, 110*, 124–129.
Velleman, D. (1992). What happens when someone acts. *Mind, 101*(403), 461–481.

Velleman, J. (2000). *The possibility of practical reason*. Oxford: Oxford University Press.
Velleman, J. (2006). *Self to self- selected essays*. Cambridge: Cambridge University Press.
Xu, K. et al. (2015). *Show, attend and tell: Neural image caption generation with visual attention.* JMLR: W&CP, France.
Zalla, T. (2007). The cognitive role of phenomenal consciousness. In M. Marraffa, M. De Caro, & F. Ferretti (Eds.), *Cartographies of the mind: Philosophy and psychology in intersection* (pp. 189–199). Dordrecht: Springer.

Danielle Swanepoel is an assistant professor of philosophy at SolBridge International School of Business in South Korea and a research associate at the University of Johannesburg, South Africa. Her current research area is primarily focused on rationality, agency, and philosophy of artificial intelligence. Her other research interests include philosophy of action, consciousness, neurophilosophy, and philosophy of mind. She is currently working on projects focusing on constitutivism, as well as exploring the possibility of agency, rationality, and morality in artificial intelligence.

Chapter 5
Consciousness: Philosophy's Great White Whale

Gerald Vision

5.1 Introduction

A notion circulating in some prominent philosophical circles is that AI resolves, or at least makes sufficiently concrete, age-old philosophical queries about the *understanding* of human minds. It is clear that computers can perform many intellectual tasks n-times faster and less fallibly than humans, even when we combine our efforts. "Artificial INTELLIGENCE" is not a misnomer. But as a surefire model for mentality in general, including that of humans, the looming question has been "Can these machines give rise to (phenomenally) conscious aspects, here CAs?" It is not in question whether we experience CAs. If computers can't do it, how do they achieve an understanding of our mental lives? So a first question here might be," Could phenomenally conscious aspects be installed in artificially constructed machines?" If the answer is affirmative, a second question arises concerning mind-body kerfuffles within philosophy: "Does this show that CAs are ultimately reducible to matter (say, brain states), thereby serving the ends of *metaphysical* materialism?" My answer to the second question is that it fails to advance the metaphysical disputes one iota. Unfortunately, I can't begin to argue for that. Those conflicts raise myriad epistemological questions going well beyond my current charge. But below I indirectly address at least one facet of my answer. Returning to the first question, I haven't a solution in hand. Toward the end of this chapter I offer some speculations about artificial machines and CAs. Unfortunately, these new AI trousers leave the older mind-body conundrums from the second question just where they were in their traditional pantaloons. But I hope it sheds some light that is central to our first question. Still, both questions raise the need for a clarification about the place of CAs in our mental economy. The bulk of this discussion is devoted to that.

G. Vision (✉)
College of Liberal Arts, Temple University, Philadelphia, PA, USA
e-mail: gvision@temple.edu

R. W. Clowes et al. (eds.), *The Mind-Technology Problem*, Studies in Brain and Mind 18, https://doi.org/10.1007/978-3-030-72644-7_5

The route our questions have taken naturally implies that humans (and no doubt other creatures) are sentient beings. From there we can inquire into the prospects for how such sentience may come about. Sentient episodes have been variously depicted. On a popular portrayal they are the "what-it-is-like" of experience, or as the undergoing of CAs. Examples abound – the taste of cinnamon, feeling tired, pains such as a toothache or of a paper cut, seeing a lilac bush, an itch, or hearing the sound of a violin, among a myriad of similar episodes. I use the term 'aspect' loosely to cover states, substances, qualities, features, events, episodes, properties, facts, or processes. For an official statement of this assumption I use

> (Ph) There are creatures, *e.g.*, humans, who experience phenomenally conscious aspects (CAs) for which there is something-it-is-like for them.

(Ph) by itself is compatible with the great swathe of traditional positions on the mind-body problem, running from Cartesian dualism to tough-minded materialism. Although, as I mention briefly later, some authors have rejected (Ph), here I want to focus on differences between theorists who accept it and offer explanations of its appearance.

Ruling out metaphysical idealism (Ph) presents us with the following choices about the timing of the appearance of CAs:

(i) They arrive in the world sometime during the development of corporeal, and otherwise nonconscious, reality, but after the onset of corporeal development.

(ii) They are equally original ingredients of reality.

For the present inquiry (ii) is qualified in two ways. First, I exempt Cartesian dualists: unlike the protagonists discussed below, their view is not designed to contain a naturalistic account of the integration of CAs with non-mental reality. Second, I also overlook an elaboration in which CAs are original only because they are forever supervenient on, grounded in, or realized in a more basic something.[1] Without such qualifications (ii) would fail to be consonant with the thrust of the case made by its adherents. Indeed, as I hope will become clear, such dependencies could be as easily accommodated within (i) with minor adjustments.

A choice between (i) and (ii) is not a usual way of laying out the options concerning phenomenal consciousness, but it sets up what I take to be a crucial difference once we have settled on (Ph) – namely, a conflict between emergentisms (i.e., theories of type (i)) and monistic panpsychisms (i.e., theories of type (ii)). They propose divergent narratives about the origins of CAs.

With respect to consciousness, emergentism is typically regarded as a dualist view. Although CAs may be grounded in a physical base, on it their non-material character is to be distinguished from that base.[f] However, from the perspective of panpsychist-leaning monism the relevant physicalism (or materialism, terms I use interchangeably for our limited purposes), which holds that CAs are identical to or strongly supervene on matter, is also a form of emergentism. The dualist variety is

[1] My thanks to Phil Atkins for bringing this to my attention.

then just a form we may call *radical emergentism.*[2] On either form of emergentism the world may begin (logically, analytically, and historically) with only physical, non-conscious features, and in the course of its development there arise CAs through sheer complexity or the right combination of its ingredients. In contrast, a monism of (ii) that accepts (Ph) – say, that inspired by Bertrand Russell (1927) – rejects that account of CA's origin. Its adherents criticize any sort of emergentism on the grounds that what arises has a character so fundamentally unlike anything that could be read from its base alone.

Emergentist's critics find it particularly unsettling that the view must construe the relation between the material and phenomenal as what David Chalmers calls "primitive".[3] He writes,"[e]lsewhere the only sort of place one finds that sort of primitive principle is in the fundamental laws of physics" (2002, p. 254). The suggestion is that introducing such an identity or grounding relation at a level other than a fundamental one, thereby precluding the connection between matter and conscious episodes from further explanation, is, (in a quip borrowed from Russell), "more akin to theft than to honest toil" (Chalmers, *Ibid*). Less colorfully, Chalmers writes that interjecting the primitive at this non-basic level is *ad hoc*, a verdict presumably sufficient to exclude it from serious consideration.

Among non-emergentists who acknowledge (Ph), only a sort of view that makes (Ph) independent is left standing. That leads fairly directly to some form panpsychism and a monism embracing both the physical and the mental. Its status rests firmly on the fact that it dispenses with those brute relations. Put otherwise, the point made in the preceding paragraph is not only an objection to (i), but with (Ph) in the books it proposes something falling under (ii). Because of this pivotal role in both defending *and understanding* our sentient life we must eventually cast a more critical eye on panpsychism. Accordingly, let us first fill in details about the nature of this replacement for (i) before seeing whether it can work.

[2] See Seager and Allen-Hermanson (2010) and Van Gulick (2001). (In Philip Goff's substantial revision of the Seagar and Allen-Hermanson entry, this dichotomy is no longer part of the exposition, but it is not repudiated, and panpsychists widely regard the options this way although their titles for the non-panpsychist alternative may vary.)

[3] And not them alone. This is a staple of non-radical emergentists of every stripe. For example, (non-monist) Colin McGinn writes "It is implausible to take these correlations as ultimate and inexplicable facts, as simply brute. And we do not want to acknowledge radical emergence for the conscious with respect to the cerebral.. .." (McGinn 1989, p. 353).

5.2 Russellian Monism

For (Ph) monists propose (a) that conscious aspects are original ingredients of reality, not resulting from its development, and (b) that those aspects are not substantively distinguishable from the physical world.[4] But, first, let us clarify what we are to count as physical.

A consensus of materialists hold that prime physicality is discovered at the lowest level of particles of a final (or modestly idealized) physics. The macroscopically physical world is built up both causally and mereologically from the behavior of fundamental particles, those at the very basis of subatomic physics. Whatever interactions occur at that level are the final arbiters of physicality. Macroscopic reality is physical only insofar as its material origins and causal powers derive from the behavior of those particles. These relationships are inscribed in the lawlike generalizations between the particles themselves, their relation to higher level physical aspects, and those between physical aspects at various levels in this hierarchical structure.

It would also be possible to launch one's materialism (or physicalism) starting from everyday specimens of material objects, almost exclusively macroscopically physical things. That would be a brand of object physicality, or, using Stoljar's (2001) handy distinction, the O-physical (ordinary, or observational) as opposed to the T-physical (found in physical theory). Contemporary physicalists typically shy away from basing their view on O-physicality because it doesn't perform the explanatory role viewed for physics. So for this discussion consider only the dominant strain of T- physicalism.

However, there is an important quirk in physics' approach to its subject-matter; it is exclusively concerned with relational or dispositional (*viz.*, dynamical) traits, not with the intrinsic natures of the properties or things possessing them. The laws of physics deal only with dispositions and relations, so-called structural features, of its subjects. Its disclosures never reach beyond those behavioral tendencies. This is not unique to physics: the same is true across the special sciences. Chemistry determining what is chemical and biology determining items in the biosphere deal only in the relational and dynamic features of chemicals, its compounds, and from organisms.

Take Boyle's law as an example. The *pressure* of a fixed amount of an ideal gas is inversely proportional to its *volume*, two properties of the gas. (Nothing about this changes by adding the Charles and Gay-Lussac elaborations.[5]) Or take Columb's Law, which tells us only about the *force* between particle charges. Even physics' fundamental forces – gravity, electromagnetism, weak and strong nuclear interactions – are understood only through their behavioral tendencies. Bottom-line

[4] (b) marks yet another difference from Cartesianism.

[5] The eventual combined law PV/T = *k*, with T as temperature in kelvins and k a constant of units of energy divided by temperature.

fermions and bosons go no farther; we distinguish them by way of their respective half and whole integer spin properties.

Different monists may describe this limitation in different ways, but in the end it amounts to much the same thing – it is the behavior of properties via their relations and dispositions, not the things themselves that have them, which science investigates. Russell (1927) characterizes these as structure and relation-number, Chalmers as structural and dynamic properties, where in both cases structure amounts to the relation of something to its surround (*viz.*, its role in a larger system) or simply logical features. Russell also describes the properties as mathematical, and the increasing mathematization of physics since his time, as Susan Schneider (2017) emphasizes, further illustrates that it is the relational, functional features that matter. As Russell writes: "A piece of matter is a logical structure composed of events; the causal laws of the events concerned, and the abstract logical properties of their spatio-temporal relations, are more or less known, but their intrinsic character is not known" (1927, p. 384).

This is inherent to science, the very heart of the enterprise. It was true in Aristotle's time, persisted throughout the middle ages, and remains science's polestar. On Aristotle's explanation of natural motion, the speed of a falling body is directly proportional to its weight and inversely proportional to the density of the medium through which it falls. Galileo took the same property (speed of fall) to be directly proportional to the square of the time it takes to fall, independently of its weight. None of that scientifically elucidates what has those properties 'intrinsically', 'categorically', or 'non-relationally'.[6]

This need not be a prescription for every scientific enterprise. For example, there is also taxonomy. But it would be a thin shadow of a scientific investigation that stopped with such classifications. The chief interest is in what those taxa are doing and how they get on the roster. Ultimately each of them is explicated in terms of its relational and dynamical properties.

These monists reject the notion that the world might be nothing more than a collection of relations. Even if there are no substances, only properties, they cannot be just relational or dispositional. An intrinsic object or property is needed to bear them. Repudiating the need for something intrinsic to pull this off implies that the whole world could be comprehensively described relationally. This may be done on a limited basis for functional theories of mental states such as beliefs and desires. But even here it is taken for granted that there are intrinsic somethings underlying these roles, even if some or all of them are non-mental. Many find a purely relational world not only implausible but incoherent.[7] Relations between *what*? *Whose* dispositions? However, past the fact that there must be intrinsics, we are not in a

[6] Questions have been raised about each of these ways of characterizing the things themselves. For present purposes as long as we agree that there are also non-relational properties/things, we can bypass disputes over the best way to explicate intrinsicality.

[7] As George Theiner and Mark Bickhard brought to my attention, this isn't a problem for process theories. But that view opens up a host of much wider issues that aren't factors in the views currently under consideration.

position to say much more about them. Following Barbara Montero's (2015) lead let's call them the 'inscrutables'. (Later we shall need to consider Chalmers' "scrutability thesis" (2012), which leads to the awkward position of having to discuss conditions for the scrutability of inscrutables, although Montero's intrinsics are inscrutable *de facto* whereas Chalmers' scrutables are *de jure*. I hope that the continued use of Montero's term for the in-themselves allows us to overcome future confusion when we must raise scrutability issues.)

Monists as such need not declare that inscrutables have the intrinsic character of conscious aspects, contain conscious aspects, or are related intimately (say, as protophenomena [Chalmers]) to CAs. But if monists are to explain (Ph), they must opt for something from that menu. If *all* the inscrutables contain one of these features, we have arrived finally at *panpsychism*, the view that consciousness is at the very foundation of the rest of the world, including the physical world as it is open to our scientific understanding. The problem monists cite for emergentism is thereby avoided by not having consciousness *arise* from anything. CAs or their bases have the status that physicalists claim for the basic entities of physics, as an original component of actuality. This version of panpsychism shares with dualistic emergentism the view that CAs are neither identical with nor have their intrinsic characters explicable in terms of – that is, are reducible to – a material base.[8] As (b) indicates, panpsychists even declare a role for the independently phenomenal in bringing about the physical events within our ken.

5.3 Panpsychic Details

To explain how monistic consciousness can be part of reality, fundamental particles must depend on or have attached to them an intrinsically conscious or protoconscious aspect. (For simplicity, I start from conscious aspects; the protoconscious makes an appearance later.) Inscrutables needn't be exclusively conscious, but they must at least include it. Whereas inscrutables that are strictly O-physical are consistent with monism, they wouldn't figure as such in the explanation envisioned for (Ph). As noted, monists tend to detail very little about the nature of the consciousness lying at the bottom of this inverted pyramid, other than stating that it is so different from *our* sentient experiences that we cannot even imagine what it must be like. It follows that the facts of consciousness go well beyond our commonplace conception of it, the latter of which embraces only the experiences of selective groups of organisms.

However, pansychism's primordial consciousness not only provides a distinctive account of the origins of higher level sentience, but it also accords CAs a role in the formation of the physical realm discoverable by science and observation.

A few preliminary points are of special interest.

[8] On the other hand, Galen Strawson (2008) holds that this makes CAs themselves material. For more on this see below.

First, it is plausible to suppose that the dependence of the scrutable world on the inscrutables is causal. If so, it must be conceptually antecedent to the brand of causation sketched by interactions in physics. On a standard physicalism causal efficacy is ultimately traceable to the action and activity of fundamental particles. That is not only the basis for all recognizable causation at subsequent levels, but it also accounts for the *understanding* of causation itself. However, if physical particles *causally* depend on the inscrutables, how could our understanding of causation derive from the workings of particle physics? On the other hand, if the inscrutables are causally inert, the point of appealing to them as a way to understand causation is nullified. One way around this would be to declare that this ultimate dependence is constitutional rather than causal. But much more must be said to develop that view. This is worth keeping in mind if only because our CA's causal dependence on the inscrutables then places in jeopardy its usefulness for panpsychism's narrative. That tells us why panpsychism is committed to describing its advantages over emergentism without appealing to this common ploy.

Next, we must qualify the significance of monism and panpsychism for the defense of option (ii). Nothing monists have said thus far rules out the possibility that there are two or more separate kinds of inscrutable, only one of which has a conscious or protoconscious aspect. In fact, that alternative would help to explain why only some macrophysical entities seem capable of displaying consciousness. Nevertheless, this would be pluralism rather than monism. It would invalidate something panpsychists take as distinctive of their view. However, that departure would not detract from the view's basic contention that conscious aspects are original explanatory ingredients of reality. Consequently, the rationale behind positing inscrutables – that something intrinsic must generate those relational features—remains intact on that revision. Panpsychism's general approach to the origins of CAs would persist even if its monistic claims were abandoned. However, while both observations affect the plausibility of any form of pansychism, I shall ignore them here. They make no difference to the points I now wish to raise.

Let us then turn to an examination of panpsychism's *bona fides*.

5.4 Polemics

A widely circulated objection to panpsychism is the combination problem. It can be found as early as in the writings of William James (1890). Panpsychists must explain how accumulating these tiny bits of consciousness can combine into a single whole to account for the consciousness with which we are familiar. Thus far no suggestions to accomplish this have been offered beyond highly speculative, indeed fanciful, ideas. Perhaps there is a yet-to-be-realized engineering technique out there. It is tempting to minimize the difficulty by treating it as nothing more than a mereological task, building up to the familiar consciousness clearly intended by (Ph), like a beach built upon many grains of sand. However, bad enough as that analogy is, focusing on it fails to appreciate an even more intractable difficulty, namely a numerically

distinct subjective episode to different possessors. This is more than accumulating the tiny to achieve the enormous. As Thomas Nagel (1974) famously noted, conscious states are inherently subjective: their 'what it is like' feature is their very essence. How tokens of inherently subjective states can be transferred from one subject to another is highly problematic. Could I experience the *numerically same* sensation that you are experiencing?

A tempting response might be that the case of different subjects with numerically identical CAs is different because the two subjects are not really distinct, but rather progressive states of oneself. But that won't work here. The initial CAs belong to tiny homunculi inside us, not to the larger individual. It is akin to taking the pain, say, of ten different people and constructing from them an enlarged pain that belongs not to each individual, but to a composite (and different) individual made up out of ten of them. This is the transference of CAs that strikes one as so troublesome for panpsychism's calculus. Thus, we have not just the conundrum of building up a robust consciousness out of unimaginably tiny distinct consciousnesses, but also of transferring an inherently subjective property to a distinct, perhaps communal, subject. But I leave that issue there to concentrate on other complications for the view.

Perhaps even more problematic, it sounds wildly implausible to ascribe any sort of consciousness to the most unstructured fundamental particles, and we are scarcely likely to overcome our incredulity to be told that those CAs are so primitive that this consciousness is radically unlike anything we experience. Quarks, leptons, bosons, or their successor particles will need to possess it. That leaves an enigma at least comparable, possibly worse, to the one with which monists charge emergentism.

David Chalmers (1996, 2015) acknowledges the point. Let us track how it strikes me that he addresses it. To alleviate the counter-intuitive attribution of CAs to particles he suggests that the inscrutables may not themselves be conscious aspects, but more primitive entities that give rise to them, either when combined in certain ways or at a certain stage of development. The inscrutables are *protophenomenal* rather than phenomenal. He calls this panprotopsychism. To avoid proliferating polysyllabisms let's continue to regard it as a variety of plain old panpsychism.

At first glance this is no solution. If the inscrutables are protophenomenal only because they are involved in the rise of the phenomenal – a view thus far barely distinguishable from the discarded emergentist solution – why isn't hydrogen protowater or a stone slab a protocathedral? If that were the end of the tale panprotopsychism would be vulnerable to this *reductio*. But, as we might expect, this is but the first step in Chalmers' proposal. Protophenomenalism's properties are specifically designed for phenomenal properties. As Chalmers explains "phenomenal properties are logically supervenient" on protophenomenal ones (1996, p. 126). And he cashes this out as the requirement "that there is an a priori entailment from truths about protophenomenal properties (perhaps along with structural properties) to truths about the phenomenal properties that they constitute" (2015, p. 260).

However, now a *reductio* threatens from a new quarter. If there are entailments, much less a priori ones, they are the products of what we think, say, or write about protophenomenal and phenomenal aspects, not features of the phenomena

themselves. It is here that the notion of an entailment has play. For convenience call these vehicles *propositions*. The difficulty with this response is that the propositions in question can be formulated in importantly different ways. It is easy to manipulate phrases to make this come out hospitable or inhospitable to an a priori entailment. This makes the possibility of a priori entailments for any subject both too easy and trivial.

An analogy illustrates the problem. The essence of hydrogen can be captured essentially by

(1) Hydrogen is an atomic element with one proton and one electron.

Now take a particular instance of water, which I name '*w*'. Chemistry has it that
(2) Water *w* is H_2O (*viz*., H–O–H).
No hint of an a priori entailment of (2) by (1) here. But suppose we enhance (1) as
(1*) Hydrogen is an atomic element with one proton and one electron, which when combined molecularly with a certain oxygen atom and a second hydrogen atom in a molecular compound, results in *w*.

(1*) still doesn't entail (2). For example, it is unclear that a single molecule suffices for *w*. But we can see in rough outline how to beef it up to an a priori entailment along the lines of something being water by something being hydrogen.[9] This is achieved simply by adding to the intrinsic description in (1) the capacity to interact with something else. By additional applications of this method we could add to (1) formulas for hydrogen bombs, ammonium, methane, hydrogen peroxide, hydrochloric acid, benzene, ethanol, and untold thousands of other compounds that contain hydrogen. Indeed, given the enormity of a universe almost totally unknown to us, a list of H^+'s dispositional properties is quite possibly incompletable. The entailment is the product of one among a multitude of descriptions of hydrogen. If (1) suffices to characterize hydrogen, any addition to it from our list is scientifically and rationally arbitrary; it trivializes the contention that we can get a priori entailments in this area. The option of framing a description containing all the hydrogen compounds would be needed, of which our current science contains only an infinitesimal sample. Is such enhancement a realistic prospect?

So let us then ask why the same isn't true for the relationship between our ways of describing protophenomenal and phenomenal aspects. Of course, a priori entailments will be attainable! How could they not be? It would be remarkable if protophenomenal aspects *could not* be formulated in ways in which the resulting proposition entailed phenomenal aspects. Could *anything* fail to be describable in similarly numerous ways?

Chalmers later proposed a way to avert this catastrophe. His *scrutability thesis* comes into play at this juncture. The thesis states that the world is comprehensible, "at least given a certain class of basic truths about the world" (2012, p. xiii). Intrinsic characters are *quiddities*, that is, "certain intrinsic properties of (microphysical) things . . . that play certain microphysical roles" (2012, p. 348n18). Moreover,

[9] Cf. Carl Hempel (1965, p. 260).

quiddities constitute whatever bears those intrinsics. Although scrutability as such is generally applicable to a host of subjects, our interest is confined to conscious or protoconscious aspects lying at the foundation of our systematic knowledge. A desideratum of scrutability is having in hand not only rigid designators, but what Chalmers calls super-rigid designators of those quiddities.[10] He then lists four concepts that we, or our idealized counterparts, need for grasping *thick* quiddities. (Below I shall skip running through the whole list. Only one, the Lewisian gambit mentioned below, is relevant to this inquiry.)

Before embarking on a quest for scrutables, we should be on guard against a potentially misleading suggestion. Listing conditions of scrutability for an X does not show that X is logically coherent. If being possible is a requirement, it does not even show that intrinsic consciousness is scrutable. Simply enumerating the conditions that must be met for anything to be scrutable could apply even were its destination inconsistent. For example, a round square has the quiddity of a radius of π with four equal sides. Of course, there is no current reason to believe that protophenomena are self-contradictory. But the point is quite general. In our case, giving legitimate conditions for scrutability does not show that the protophenomenal gets a pass out of the inscrutables, even for our idealized counterparts. This may be brought out by Chalmers's response to David Lewis's (2009) argument for "Ramseyan Humility".

As should become clear, this Lewisian stratagem simply replicates a problem illustrated earlier. To sum up the argument, Lewis claims that with respect to Ramsey sentences for fundamental properties – say,

$$(?x)(F_1x \,\&\, F_2x \,\&\, F_3x, \&\, ?\&\, F_ix)$$

– we are in no position to pick out the predicates/properties in those conjuncts that are more than role occupiers. Even if that can be overcome, equally permissible different selections of those subsets renders arbitrary the choice of any one of them as an source's distinctive quiddity. That was the point put more bluntly in my hydrogen example. Chalmers then proposes the following fix:

> replace the existentially quantified claims that there are properties that play the roles by claims to the effect that such-and-such properties play the roles, where 'such-and-such' expresses quiddistic concepts of those properties (2012, p. 351).

In other words, instantiate all or part of the quantified formulas with singular sentences in which quiddistic designators of the particles are the instantiators. We may construe this in either of two ways.

First, suppose the quiddity is intended as the whole conjunction of properties. If Chalmers' notion of quiddity allows this, I do not see a way of denying that any

[10] This notion comes out of Chalmers's two dimensional semantics, demanding that the designator be rigid both epistemically and metaphysically. We need not say more about that refinement. I mention it here simply to bring out the epistemic – that is, *a priori* – part of the appeal to a priori entailment. It plays no further role in the points under consideration.

instantiated Ramsey sentence displays the quiddity of this instance. But it runs counter to what monists, including Chalmers, state elsewhere. For one thing it would include all the relational and dispositional properties of the subject. But Chalmers defines a quiddity as "[t]he *categorical basis for* microphysical dispositions" (2012, p. 473, my emphasis). He remarks that his earlier a priori entailers could be supplemented by structural properties. But the same doesn't hold for quiddities. Quiddities were introduced to distinguish what the scrutable *is* from what it *does*, thereby excluding merely structural properties.

A less plausible interpretation would be to choose from among the predicates those that pick out the quiddity. However, unless prior to the construction of a Ramsey sentence we knew which those were, why is any choice non-arbitrary?

But there is even a more basic shortcoming lurking in this strategy. On the Lewisian model the everyday or scientific properties that allow us to pick out the unique predicates for which we are searching doesn't disclose for the generalized subject *what* it is. Yes, the something-or-other uniquely has these properties, so we have at hand the means to avoid confusing it with something else. But that does not yet disclose *what* has been located. It is like a blip on a radar screen that uniquely locates an object without disclosing what has been located. It individuates something without identifying it: and *individuation is not yet identity*. The former is neither an adequate substitute for the latter nor is it an assured recipe for finding it. For example, if the police know that the suspect is the only person in the house, it does not follow that they know *who* the suspect is. Suppose Chalmers assumes that Ramsey sentences perform the work outlined for his inaugural characterization of Laplace's demon. Let us say that this superior being is able to deduce any particular truth from a knowledge of all the basic truths. But if the knowledge Laplace's demon has is only what microphysics yields, then on monist premises the demon could yield only future locations and dynamic properties. That remarkable feat would get us no closer to the intrinsic natures of the possessors of those properties.

To sum up, the accumulated difficulties with panpsychism arising from the combination problem, plus those arising from its initial incredulity – viz., assigning consciousness to fundamental particles – remain. Nor has the incredulity problem been minimized by these heroic efforts, from protophenomena to quiddities, to solve it. It would be an understatement to say that the panpsychist account of the conscious is *in extremis*.[11] Where does that leave us?

5.5 Emergentism Redux

We have restricted our investigations to views satisfying (Ph) – that is, those that attempt to account for a subject's conscious aspects. Suppose that the various fixes to panpsychism we have reviewed don't relieve it from its serious difficulties. The

[11] None of this speaks against (or for) monist proposals that are not attempts to support (ii).

only alternative is (i), one or another form of emergentism. I have not mounted any initial case for emergentism, or even tried to fill in details beyond the general sketch given earlier.[12] It would be a Herculean task to argue *de novo* for so sweeping a postulate here. Rather I have thus far set out to expose difficulties in choosing (ii) over (i). Of course, there is no guarantee that if emergentism was subjected to a comparable level of scrutiny, it wouldn't expose problems as grave as those just discovered for more rigorous forms of panpsychism. However, given that emergentism runs the gamut, from current forms of physicalism to those of dualism, it seems we can be confident that if there were any such absolute disqualifiers they are likely to have been exposed by panpsychists or others who have contributed to the copious literature on consciousness. So, rather than replaying the whole past debate over emergentism, I shall simply comment on the central point of departure that the general run of panpsychists, including Chalmers, see as dividing their preferred view from emergentism—namely, the latter's need for a primitive or brute relation to phenomenal consciousness at an intermediate level. Put otherwise, the view under attack is the introduction of a kind of fundamental occlusion at an evolving point in the material world. Even here, there is no space to do adequate justice to the limited topic; I confine myself to a small selection of noteworthy points.

First, unlike Chalmers's restriction of the *primitive* to an absolute lowest floor, and its formulas at which the basic laws of science reside, emergentist primitiveness is a *relation* between two things, matter and CAs. This relational notion is a different phenomenon from the bottoming out of ultimates, in which what is primitive is wholly unsupported by anything else. It is a different way to carve up the landscape, one for which bottoming out may be a poor basis for comparison. Indeed, the use of *primitive*, carries the misleading suggestion of reaching the terminus in a hierarchy. The notion of a brute "relation" carries no such suggestion and is thus less misleading.[13] We are discussing a relation between things for which a certain favored kind of account providing more substantial detail about the connection between two well-understood terms is not in the offing. It is not the end-point of an all-embracing hierarchy, as were, say, the simple ideas of Locke (1700, II, ii, §§1–2) or the impressions of Hume (1739, I, I, §1).[14]

Now it may be that when we reach fundamental particles, nothing remains to be explained. For that reason some have been tempted to regard the fundamental laws of nature as necessary in a sense stronger than natural necessity. It would not be idle to suspect that this is no more than a gratuitous compliment to being basic. But whatever the grounds for that status, necessity here suspiciously resembles the old Principle of Sufficient Reason: if something's *raison d'etre* does not belong to its relation to other things, the reason for its existence must be found in the thing itself. From there it may be deemed a short step to making that thing a self-explanatory,

[12] However, in this section's last two paragraphs I shall comment briefly on the respective debating positions of emergentism's physicalist and dualist forms.

[13] This term is employed by the hesitant panpsychist William Seager (1995).

[14] Widespread disparagement of brute relations is discussed in Vision (2018). Much of the discussion of the next few pages summarizes that more detailed treatment.

necessary being. Without going into detail, a violation of a principle of that flavor is what panpsychists would need to charge emergentists with if they are to defend their dismissal of brute relations as beyond any credibility. However, the Principle of Sufficient Reason isn't self-evident and indeed is controversial. We should demand ample further support.

Still, on the other side of that coin this may seem to be contingency without any explanation. What can be said in the defense of that prospect?

Although I would prefer to avoid navigating the minefield of theories of explanation, we must at least poke around its edges. For the present objection rests on a certain widely accepted condition for something to count as an explanation. This is less than settled doctrine; it is at best unclear that citing a brute relation is *not* an explanation. It's an asset for explanation to reflect ontology, and if a relation is brute, why doesn't stating so count as *explaining* it? That we should *want* more is understandable; that we should *demand* it needs a better rationale. And its support should be more than the adoption of a theory of explanation that excludes brute relations (or contingent termini) for no compelling reason. A grimace isn't a serious objection.

Finally, it may appear that proponents of this objection to emergentism have, as Bertrand Russell acidly noted (1918, p. 179), simply relegated the unfamiliar to the unintelligible. If that is the gravamen of the charge it is certainly vulnerable. However, it is unclear that the opponents have established even the unfamiliarity of brute relations. For example, consider identity statements involving rigid designators. They are necessary truths and at least occasionally look as if they are brute. No doubt, in certain instances further reasons may show that an identity is more than brute; the continuous spatiotemporal coincidence of Voltaire and François-Marie Arouet supports numerical identity. But what more can be said about the identity of gold and the element with atomic number 79? That it is a necessary truth explains *where* it holds (*viz*., everywhere possible), not *why* it holds. A counterclaim that its necessity is conventional is confronted with the labor that went into its discovery. Moreover, conventions can be changed, re-stipulated; it would be absurd to suppose this is in the offing for gold (*qua* rigid designator). Such cases may provide a fertile source for brute relations.

Also, what could critics of brute relations have to say about the dependence of chromatic color on a non-chromatic reality? The dependence of colors on the colorless applies not only to color realists, but also to secondary quality and error theorists. For simplicity consider only the colors of objects (not that of spectral light). If colors inhere in their objects, they nevertheless supervene on what is itself chromatically absent (say, wavelengths between roughly 380 and 760 nms). On secondary quality theories colors arise from an interaction between perceivers and the material world. But that only relocates where their brute connection occurs. What more is there to lively colors of objects arising from a relation between a *blah* perceived world and its *blah* brain matter, any more than they can arise, say, from the selective partial absorption of wavelengths? And for error theories colorful experiences must have their basis in something. Even hallucinations and delusions are caused. A brute relation somewhere is inevitable.

I conclude that the present critique of emergentism is not the mortal blow some may have envisioned. Of course, none of this shows that a broadly characterized emergentism of either a materialist or a dualist flavor succeeds. We are far from having exhausted the objections to them. But it does speak to the argument(s) panpsychists use to highlight their preference over any other way to account for CAs.

What of those who reject (Ph) as it stands? Here I must be very brief. An eliminativist, for example, might suggest simply to drop (Ph), or, a reductionist might reclassify conscious aspects as abilities for action rather than as episodes. However, both views are not only radically counterintuitive, but also imply that we are massively and inescapably deluded. They start from a patently implausible supposition. That is not because appearances are everywhere inviolate. Some wide-spread, stubborn appearances have proved delusive. The two lines in a Müller-Lyer diagram persist in appearing of different lengths, even after we know they are the same length. But we also have an explanation of the reality behind this deceptive appearance. However, what possible explanation could there be to fall back on for the massive delusion that we undergo experiential episodes? Where is the gap between appearance and reality to account for it?

But even if we have yet to find a compelling reason to reject either conscious episodes or a material base from which they arise, I have done nothing to distinguish, much less support, the dualist version I believe (if you'll pardon the expression) emerges as a robust option among the remaining contenders. Of course, one among various grounds for suspecting it is that it has become an orthodoxy among a stable minority of thinkers, and orthodoxies of all stripes have suffered dismal fates. Nevertheless, dualistic emergentism still seems to have a pulse, which I believe is the best that can be said for it.

Physicalists, on the other hand, are likely to state their version with either of the following claims: (a) whatever is supervenient on (or grounded in) the physical is *eo ipso* physical or (b) conscious episodes are (type or token) identical to their physical bases. These are daunting options. (a) should strike us as a seriously weakened version of physicalism (Kim (2000) calls it "minimal physicalism"). Of course, there are different types of supervenience, of varying strengths. Still what supervenes on something, at least without further explanation, no more needs to take on the character of its base than a butterfly needs to be a caterpillar. In fact, it is difficult for me to see how this requirement alone allows the view to differ in anything more than name from either property dualism or, perhaps, the neutral monism of property possessors. As for (b), the issues raging around that hotly debated view are too complex to broach here. Still it is worth noting that monists and panpsychists seem to have happened upon an elusive something, although it was not their aim, in raising doubts about how we can understand it. Their doubts rest on the *aperçu* that we don't know what to make of a strictly physical something (as we understand that) that has an angle, aspect, or side but seems to lack anything that in isolation suggests its physicality. This is radically unlike the usual situation in which we can be surprised about the identity of something encountered under different circumstances, but in which there is no basis for puzzlement about each encounter

being of something with all the marks of physicality.[15] Here we have an additional puzzle. We must be given some grounds to understand this doubling of surprises. It is of an entirely different order from other cases of the identities of physical particulars, and it is not obvious that such misgivings can be overcome by theoretical considerations alone. In any event, such considerations seem to have whatever purchase they can muster at the level of types, and are seriously handicapped when restricted to the level of a more popular token materialism.

The final score? Even if an artificially designed machine is proposed as the foundation, say of supervenience, for CAs, all the old issues dividing dualists from materialists remain. If we are not eliminativists, nothing is added to the debate when the material counterpart is silicon rather than carbon. And, if we are eliminativists, we didn't need it added to AI's problem solving skills to begin with.

If panpsychism were our starting position it is even more obvious that materialism needn't be in the cards. Mildly oversimplifying, panpsychists maintain that every scrap of anything is already endowed with a primitive form of consciousness (or protoconsciousness). Advances in AI can't improve on that. Indeed, on that view it is difficult to see what it could mean for scientists to build into a synthetic product an original mature consciousness, one that was not an accumulation of its already present extant bits. Nothing built into those systems changes, or improves its position with respect to consciousness. Although there are eccentric characterizations of materialism so that any foundation will be labeled "material",[16] *prima facie* this looks like an immaterialist view.

Is there then a place for considerations of AI in any of this?

5.6 Back to Prosthetic Consciousness

Computer technology is rapidly catching up to Hollywood. What was once mere science fiction may now be just over AI's horizon, or has already happened in top secret government laboratories. However, what was probably obvious to many all along, a discussion of the place for CAs in artificial intelligence is in one sense off-topic. AI's serious pursuit is a super-intelligence. The addition of phenomenal consciousness may be a central concern for those investigating whether artificial intelligence is of guidance in understanding human minds. But this is a detour from the main issue raised by AI and mind. Of course, it has been held by some that phenomenal consciousness is a necessary condition for mentality of any kind. Thus, if AI can lack it, "intelligence" may be misnomer for its successes. But under those

[15] Nothing novel here. In a celebrated passage John Tyndall wrote "Granted that a definite thought and a definite molecular action in the brain occur simultaneously: we do not possess the intellectual organ, nor apparently any rudiment of the organ, which would enable us to pass, by a process of reasoning, from one to the other" (1897, p. 87).

[16] Cf. Chomsky's (1968, p. 98) prescient observation that we call any explanation "physical" just because it has become established in the domain. For a move of that sort see, e.g., Strawson (2008).

circumstances this is merely a terminological issue, not the deep insight into mental life that philosophers seek. Any remarkable successes of artificial devices at the level of human problem solving would stand, however we denominated them. Mock turtle soup would taste the same. But, even if consciousness is but a side-issue for advances in AI, we can return to our initial question: "Could phenomenally conscious aspects arise in artificially constructed machines?"[17] Put otherwise, is there something so distinctive about the traditional way of creating sentient beings that rules out CAs for the product of an asexual, planned intelligent process?

One way to begin to test artificial consciousness might be to begin with a human replaced with artificial parts. We do this now with lenses in cataract surgery, and work proceeds, even if with only fitful success, on a number of artificial organs.[18] But let us imagine an even more radical case. A patient with a deteriorating condition in her central nervous system – call her 'Mary' – has her neurons very slowly replaced by (futuristic) miniature computer chips, the size of a neuron, until a certain large percentage of her nerve cells (51%?) are artifacts. (I'm ignoring the practical difficulty of achieving this within a single lifespan.) Suppose that at various stages her bodily and verbal behavior seems much as it was before the latest implant. Post-operatively in the hospitable she may behave as if she is undergoing discomfort, saying(?) she is thirsty, has a headache, and so on. It would seem to be implausible to disbelieve her after an early replacement. At what stage of continuous tiny replacements would it be more realistic to have doubts about her being sentient? If she continues to exhibit her initial characteristic behavior, why doubt that she feels what she continues to seem to feel or seems to claim that she feels?

The result of the procedure being imagined is certainly not conclusive. Were we entitled to conclude that her behavior would remain this transparently similar through even the earliest of these replacements? First person judgments might seen more compelling. Mary should *know* if things have changed. But has she(?) retained this ability of true comparison or has she become delusional about her earlier self? Can we even call what she *now* comes out with a belief? And, whose belief is it, Mary's or Intel-Mary's? Finally, do her biological origins matter, disqualifying us to generalize from her case to those in which a silicon concoction is created from non-sapient materials?

In addition we don't yet know whether only certain materials are capable of constituting a conscious state. For example, the carbon bond that has gives rise to organisms is being increasingly studied to replace silicon for enhanced computing

[17] Turing (1950) may be suggesting either that behavior delivers the answer or that a less ambitious question about behavior should replace it with the only real (or intelligible) one. (I take the original question to be kosher, but don't try to answer it behaviorally.) This doesn't prevent Turing from hypothesizing that conscious experience is only a matter of storage capacity, which would make him an emergentist.

[18] Given the trove, we may ask why those who have undergone cataract surgery, replacing their organic lens with a plastic one, are having genuine, rather than ersatz visual experiences. Given trove of empirical data at hand, why has this sort of case been near-universally ignored in these debates?

power. Susan Schneider has pointed out that carbon forms stable double-bonds more easily than silicon; the latter typically forms only single bonds. One speculation (it's no more than that) is that the greater strength of carbon bonds may matter for consciousness. But perhaps this limitation in silicon can be overcome by the development of neuronal based computing. On the other hand, *if* it is necessary to use, say, neuromorphic chips for consciousness, we might wonder whether the product is sufficiently distinct from the organic realm to be regarded as postbiological. Although unnaturally selected, perhaps it looks enough like an organism even if a distinct species. A brilliant achievement, but is it in principle different from synthetic bacteria that biologists now create?

I presume that consciousness and self-consciousness are almost always regarded as the single most important feature of our selves. And having conscious states is a major factor for extending ethical concerns to any group of creatures. Despite this, consciousness may be an impediment to the operation of a super-intelligence. Thus, if conscious states do attach to super-intelligences, they might only be exaptations that those intelligences can't shed, just as biology's architectural limitations prevent us from evolving wheels instead of legs, though wheels might be more useful for moving from place to place.

Like questions about whether the whole can be reduced to its parts, these uncertainties would remain after our production of super-intelligences, whether or not those creations will allow us inferior intellects to stay around to ponder them.

These are but a few of the random thoughts I can muster about the connection between consciousness as we know it and computers still on the drawing board. None of them strike me as advancing the materialism issue one whit, for reasons summarized in the last section. And here I must leave that issue in the inconclusive state from which we started.

References

Chalmers, D. (1996). *The conscious mind*. Oxford: Oxford University Press.

Chalmers, D. J. (2002). Consciousness and its place in nature. In *Philosophy of mind*. Oxford: Oxford University Press.

Chalmers, D. (2012). *Constructing the world*. Oxford: Oxford University Press.

Chalmers, D. (2015). Panpsychism and panprotpsychism. In Alter, T. and Nagasawa, Y. (Eds.), *Consciousness in the physical world: Perspectives on Russellian monism*. Oxford: Oxford University Press.

Chomsky, N. (1968). *Language and mind (enlarged edition)*. New York: Harcourt Brace Jovanovich.

Hempel, C. G. (1965). Studies in the logic of explanation. In *Aspects of scientific explanation*. New York: The Free Press.

Hume, D. (1739). *A treatise of human nature*.

James, W. (1890). *The principles of psychology* (Vol. 1). Henry Holt.

Kim, J. (2000). *Mind in a physical world*. Cambridge, MA: MIT Press.

Lewis, D. (2009). Ramseyan humility. In D. Braddon-Mictchell & R. Nola (Eds.), *Conceptual analysis and philosophical naturalism*. Cambridge, MA: MIT Press.

Locke, J. (1700). *An essay concerning human understanding* (4th ed.).
McGinn, C. (1989). Can we solve the mind-body problem? *Mind, 98*, 349–366.
Montero, B. G. (2015). Russellian physicalism. In Alter, T. and Nagasawa, Y. (Eds.), *Consciousness in the physical world: Perspectives on Russellian monism*. Oxford: Oxford University Press.
Nagel, T. (1974). What it is like to be a bat. *The Philosophical Review, 83*, 435–450.
Russell, B. (1918). On the notion of cause. In *Mysticism and logic* (pp. 171–196). London: George Allen & Unwin.
Russell, B. (1927). *The analysis of matter*. London: Kegan Paul, Trench, Trubner.
Schneider, S. (2017). Idealism or something near enough. In K. Pearce & T. Goldschmidt (Eds.), *Idealism*. Oxford: Oxford University Press.
Seager, W. (1995). Consciousness, information and panpsychism. *Journal of Consciousness Studies, 2*, 272–288.
Seager, W., & Allen-Hermanson, S. (2010). Panpsychism. In *Stanford encyclopedia of philosophy*. http://plato.stanford.edu/entries/panpsychism/.
Stoljar, D. (2001). Two conceptions of the physical. *Philosophy and Phenomenological Research, 62*, 253–281.
Strawson, G. (2008). *Real materialism and other essays*. Oxford: Oxford University Press.
Turing, A. (1950). Computing machinery and intelligence. *Mind, 59*, 433–460.
Tyndall, J. (1897). *Fragments of science,: A series of detached essays, addresses, and reviews* (Vol. 2). New York: D. Appleton.
Van Gulick, R. (2001). Reduction, emergence and other recent options on the mind-body problem. *Journal of Consciousness Studies, 8*(9–10), 1–34.
Vision, G. (2018). The provenance of consciousness. In E. Vitaliadis & C. Mekos (Eds.), *Brute facts* (pp. 155–176). Oxford: Oxford University Press.

Gerald Vision is professor of philosophy at Temple University (Philadelphia, PA, USA). He has published a number of books and articles on metaphysics, epistemology, and the philosophy of mind.

Part II
The Metaphysical and Technological Presuppositions of Mind-Uploading

Chapter 6
The Myth of Mind Uploading

Gualtiero Piccinini

6.1 Mind Uploading

If you wanted immortality and someone offered to upload your mind to an artificial computer to achieve immortality, should you sign up? There are real companies offering to freeze your brain so that someday they will upload your mind to a computer. It's worth thinking this through.

Some argue that mind uploading will happen relatively soon; others are skeptical that it's possible at all.[1] The main points of contention are whether an uploaded mind would be conscious and whether it would be numerically identical to the mind being uploaded. The skeptics argue that an uploaded mind would not be conscious and would not be identical to the original mind; proponents of digital immortality disagree. Many good points have been made but the debate is inconclusive. Much about mind uploading remains obscure.

In this paper, I will try to shed light on mind uploading by drawing on some recent work on the nature of physical computation and the metaphysics of mind. I will defend three main conclusions. First, uploading a mind to a computer is practically unfeasible and is likely to remain so forever. Second, even if we managed to

This paper descends from a talk presented at Minds, Selves and 21st Century Technology, in Lisbon, Portugal, June 2016. Thanks to Susan Schneider for inviting me and to the audience for feedback. Thanks to Peter Boltuc for inviting me to present on this topic at the 2018 Central APA, to Julie Yoo for her comments, and to the audience for feedback. Thanks to Neal Anderson, Carl Craver, Corey Maley, Jack Mallah, and an anonymous referee for discussion and comments.

[1] The essays collected in Blackford and Broderick 2014 are a good entry in this debate. Other useful references include Sandberg and Bostrom 2008, Hauskeller 2012, and Hopkins 2012.

G. Piccinini (✉)
Philosophy Department, University of Missouri - St. Louis, St. Louis, MO, USA
e-mail: piccininig@umsl.edu

R. W. Clowes et al. (eds.), *The Mind-Technology Problem*, Studies in Brain and Mind 18, https://doi.org/10.1007/978-3-030-72644-7_6

upload something like a mind to a computer, the uploaded "mind" would be unlikely to be conscious. Third, even if an uploaded mind were conscious, the original conscious mind would not survive the uploading process. Therefore, we will not upload *our* minds to computers and, if we manage to upload anything analogous to a mind to a computer, it will be rather different from *our* mind.

Before we start, we need clarity on what mind uploading amounts to. The phrase 'mind uploading' conjures up the image of uploading digital software to an ordinary computer. Most of us are familiar with uploading pictures, documents, programs, and other files. Roughly speaking, uploading software to a computer consists of setting the values of some of the computer's memory cells to match a pattern that we wish to replicate and that the computer can manipulate. If our mind were a piece of computer software, the idea of uploading it onto an artificial computer would at least make sense. Conveniently, there is a view about the metaphysics of mind that fits this idea well.

Computational functionalism maintains that the mind is the software of the brain. There are strong and weak versions (Piccinini 2020, Chap. 14). Strong computational functionalism says that the mind is the software of the brain in precisely the sense in which the programs that run on our artificial computers are their software. If this were correct, it would still be extraordinarily difficult to upload our mind to a computer, because we'd have to correctly identify the neural patterns that constitute our mind. But we could try.

An important caveat applies. Just because you can upload a piece of software, it doesn't follow that uploading gives that piece of software *immortality* in the sense desired by most people who desire immortality. Immortality is infinite *persistence*, and how a person persists is different from how software persists by being uploaded to a computer. A person is an unrepeatable particular: they persist insofar as they survive as an individual that is distinct from other individuals; they do not persist (qua individual) by being copied and multiply instantiated. In contrast, uploading a piece of software means making one or more *copies* of that piece of software. To wit, you can upload the same piece of software to as many computers as you like and still retain the original copy that you used for the uploads. The original copy continues to persist as an individual after it's uploaded. Therefore, the original copy cannot persist, qua unrepeatable particular, through its other copies. In light of this, if the mind were a piece of digital software, uploading it to a computer would be a way to copy it, not a way for it to survive.

None of this matters much, because virtually no one believes the strong version of computational functionalism—not even proponents of digital immortality. Virtually no one believes it because there is no evidence for it. Instead, what proponents of digital immortality typically presuppose is something in the neighborhood of *weak* computational functionalism. Weak computational functionalism says that the mind is the software of the brain in a looser sense, which does not vindicate uploading the mind in the way we upload ordinary software. More precisely, weak computational functionalism says that the mind is the *computational organization* of the brain. Computational organization may be constituted not by a piece of

ordinary digital software but by the structure and functions of the neural networks that make up a brain (Piccinini 2020, Chap. 14).

If the mind is not literally a piece of software, what does it mean to upload a mind to a computer? Proponents of digital immortality have two suggestions. The first suggestion is to build a detailed computational simulation of an individual brain within an artificial digital computer. Building a brain simulation bears little resemblance to the ordinary process of software uploading; calling it 'mind uploading' is rather misleading. I will refer to it as *brain simulation*.

The second suggestion is to progressively replace parts of a brain with prostheses until a plurality of prostheses replaces the whole brain. Replacing brain parts with prostheses bears no resemblance at all to software uploading, because there's not even a separate computer to which the mind is putatively uploaded. So, calling this process 'mind uploading' is a misnomer. I will refer to it as *brain replacement*.[2]

Before beginning in earnest, I should clarify how I use the terms 'mind' and 'consciousness'. I assume that a mind consists in cognitive processes, some conscious and some unconscious. By 'conscious' I mean *phenomenally* conscious: a state of mind such that a subjective experience is felt. In other words, when a mind is conscious, there is something it is like to be it. I also assume that human beings are conscious at least some of the time, and that consciousness is essential enough to the human mind that surviving without consciousness is either an oxymoron or worthless.

6.2 Brain Simulation

Simulating brains is one aim of computational neuroscience. But computational neuroscientists simulate aspects of generic brains for specific modeling and explanatory purposes. They are not in the business of simulating individual brains down to their personality.

The kind of brain simulation needed for digital immortality is rather different from ordinary computational neuroscience models. It involves simulating enough details about an individual brain to represent the precise way in which that individual fulfills their cognitive functions, including their emotions and the idiosyncrasies of their mental life and character. In addition, the simulation must be able to evolve over time, in the same way that an ordinary person evolves over time as a function

[2] A third suggestion sometimes discussed in the literature on digital immortality is to record someone's experiences in a digital medium and use them to create a partial computer simulation of their mental life (Smart this volume). A fourth suggestion is to extend someone's mind into digital technology and slowly but progressively merge with digital technology until their mind continues solely in digital form (Clowes and Gärtner this volume). These proposals face analogous challenges to those I discuss in this paper. Making such challenges explicit is left to future work (cf. also Schneider and Corabi this volume).

of both their mental life and new experiences. I will refer to the relation between this kind of simulation and the original brain as *precise functional equivalence*.

Before addressing the feasibility of precise functionally equivalent brain simulation, I need to clarify what (computational) simulation is. Simulation is the representation of a target system and its dynamical evolution by a computational model. There are two notions of simulation—it's critical that we distinguish them.

The first notion of simulation is the simulation of a wholly digital computation by another wholly digital computation. By wholly digital computation, I mean a computation performed by a system whose internal state and input can be represented exactly and exhaustively as a string of digits and whose steps are discrete transitions between strings of digits. Barring hypercomputation (Piccinini 2015), and provided that a wholly digital system's initial state and internal dynamics are known precisely, a wholly digital system can be simulated exactly in a precise mathematical sense. That is, both the initial state and input to a digital computation can be precisely encoded in the state of a computational model and the model can construct successive representations of the state evolution of the target system until the model produces an exact encoding of the target system's output. The model and its target compute the same mathematical function defined over strings of digits.

If the brain were a wholly digital computing system, this notion of simulation would apply to it. Not that in practice it would be easy to build a precise functionally equivalent simulation of a wholly digital system. It would be practically unfeasible for reasons analogous to the reasons to be canvassed presently. But it would be possible in principle. Alas, neural dynamics are not digital, and there is good evidence that neural computations manipulate non-digital vehicles at least in the general case (Piccinini 2020, Chap. 13). Because of this, as astute supporters of digital immortality know, exact simulation of a wholly digital computation by another digital computation is not an option when it comes to neural systems.

The second notion of simulation is the simulation of an ordinary physical system by a digital computation. In this case, the input to the computational model encodes an approximation of the target system's initial condition together with a mathematical model of the target system's dynamics—typically, a system of differential equations. The model then relies on numerical methods to compute subsequent states of the target system to a good degree of approximation. This is the kind of simulation employed by computational neuroscientists. This is the kind of simulation that is relevant to digital immortality.

Three questions arise:

1. Is it feasible to build a brain simulation that is precisely functionally equivalent to an individual human brain?
2. Would the simulation be conscious?
3. Would the mind of the simulated brain survive through the simulation?

I will address questions (2) and (3) in subsequent sections. Here I focus on (1).

In order to focus on (1) I assume, along with proponents of digital immortality, that weak computational functionalism holds. More precisely, I assume that cognition is reducible to computation and information processing and that either

consciousness is reducible to cognition—and hence to computation and information processing—or consciousness is irrelevant to cognition. I will question this assumption later, but I need it in place for now. If this assumption were wrong, cognition would require consciousness and consciousness would not reduce to computation and information processing. This would rule out the kind of brain simulation that is needed for digital immortality.

For the same reason, I also assume that the coupling of the mind to a biological body and physical environment either makes no difference to the possibility of simulating a mind or can be simulated to the necessary degree. Some may wish to question this assumption, but I need it in place to discuss more pressing obstacles.

Even with these assumptions in place, it is unlikely that we will ever be in a position to simulate an individual brain with precise functional equivalence, which is a necessary condition to make talk of digital immortality worth taking seriously. To see why, consider the closest thing we have to a method for constructing a detailed brain simulation.

The method consists in freezing a brain, cutting it into thin slices, measuring the positions of individual neurons, their kind, and their mutual connections, and then using state-of-the-art mathematical models to simulate the individual neurons and their interconnections. There are computational neuroscientists who are involved in smaller versions of this sort of project, such as simulating a portion of rat cortex (Markram 2006; Markram et al. 2015; see also Izhikevich and Edelman 2008; Eliasmith et al. 2012; Eliasmith and Trujillo 2014). If this method were used to simulate an individual human brain, several challenges arise.

First challenge: *loss of information about dynamics*. Freezing a brain annihilates its dynamical properties. While some dynamical properties can be inferred by looking at brain structure in combination with general knowledge of brain dynamics, plenty of dynamical aspects cannot be inferred from brain structure alone. A beautiful demonstration of this is the classic work by Eve Marder and collaborators on the crustacean stomatogastric ganglion, a structure composed of 30 neurons whose wiring is well established (e.g., Hamood and Marder 2014; Marder 2012; Marder et al. 2017). Marder and colleagues have shown that there is wide individual variation between the dynamics of stomatogastric ganglions from different animals, even though they are wired in the same way. In addition, the same circuit can behave very differently, and fulfill different functions, depending on several neuromodulators. Some of this variability makes little functional difference, because the crustacean stomatogastric ganglion can fulfill its functions under many different circumstances. But some of this variability is functionally significant—neuromodulation must fit within certain values for the circuit to work properly. The important point is that neural dynamics vary from individual to individual and from function to function and cannot be inferred solely from neural circuitry. Needless to say, the same point applies in spades to the human brain and its hundred billion neurons.

Simulating an individual brain closely enough to preserve its individual mind requires simulating its individual dynamics. In a frozen brain, any dynamical properties that are idiosyncratic to the individual brain being simulated are lost.

Therefore, simulating an individual brain after it's been frozen is unlikely to result in anything especially close to the functioning of the original.

Someone might object that we can supplement this method with measurements of brain dynamics that can be made before freezing the brain. For instance, we could subject a brain to intensive scanning using fMRI and other imaging methods to identify aspects of its dynamics. But current imaging methods are too coarse-grained to yield enough information to build a detailed model of individual neuronal dynamics. More fine-grained methods, such as recording from individual neurons, are still very partial in the information they yield, too invasive to be used on human beings, and too impractical to be used on a *whole* human brain. Therefore, we are unlikely to gather enough information about the dynamics of individual human brains to be in a position to simulate them very closely.

Second challenge: *measurement errors*. Categorizing a hundred billion neurons as well as measuring their positions and their hundreds of trillions of connections comes with considerable margin of error. Many neurons and connections will be missed, others misclassified, and the measurements of their positions will be approximate. Even if the process was unrealistically accurate, such as 99.99% accurate, that would still leave ten million neurons and tens of billions of connections in error. Therefore, any simulation based on these data will not match the original very precisely.

More generally, any measurement process measures only some variables not others, has a margin of error, and interferes with the measured system in some ways. The more information we seek about a system (and the more fine-grained the information we seek), the more we are likely to interfere with the system while we gather information. Therefore, there is a trade-off between gathering more information and leaving the system undisturbed. Gathering the information needed to simulate a portion of a brain is unlikely to be possible without causing extensive brain damage and altering the brain to a point that makes the person it constitutes unrecognizable.

In addition, constructing a model from raw data requires interpreting the raw data to reconstruct structure types, connections between them, and their dynamics. Given the complexity of the system, errors will occur in interpreting the raw data and extrapolating from them. The result will not be an exact replica but an approximate description of a system somewhat similar to the original system. Depending on how much we know about the structure and dynamics of the system and how much data we collect about the structure and dynamics of that specific system, the model may describe a structure and dynamics that is somewhat similar to that of the original system. It will not match the exact structure and dynamics of the original.

Third challenge: *ignorance about complex systems*. Our knowledge of brain structure and dynamics is limited. Neuroanatomy, neurophysiology, and computational neuroscience have made great advances, but there is plenty we still don't know. Even the most detailed and accurate models capture only (some of) what we know. As scientific knowledge is incremental, we expect to know more in the future. It is unlikely, however, that it will ever be either practically or economically feasible to obtain exact and exhaustive knowledge of systems as complex as the human brain. We should not expect to ever know everything there is to know about how the

brain works. Therefore, we should not expect to know everything there is to know about how to build a precise functionally equivalent simulation of an individual human brain, or even to have models that provide an exact match of what we do know.

Measuring what it takes to simulate a brain requires that we know what to look for. There is no good reason to expect that we'll ever know enough about something as complex as the human brain to such a degree of precision that, if we could collect all the needed data without damaging a particular nervous system, we would then be able to simulate the original nervous system closely enough that we would judge it to be precisely functionally equivalent to the original brain. This is particularly true because human brain function, let alone individual personality, is not the kind of thing that we can freely tinker with in order to understand it, for ethical reasons.

To make matters even more challenging, a striking feature of complex system such as living systems is that their (dynamical) organization allows them to stay within a narrow portion of the phase space of systems made of ensembles of their components; such dynamical organization cannot be reconstructed just by knowing which components are where; if we don't know the essential organizational and dynamical properties, we won't be able to reproduce the relevant phenomena.

Fourth challenge: *ignorance about human brains*. Our knowledge of how the mind arises from brain structure and dynamics is even more limited than our knowledge of brain circuits in general. Neurophysiologists record from rat and monkey brains, but ethics prevents them from doing the same to people. As a result, we know especially little about the neural basis of higher cognitive functions that are uniquely human, such as any aspects of cognition that rely on natural language. Therefore, we have no detailed models of how the brain gives rise to *human* cognition. Lacking a general model, we have little basis on which to build a simulation of how an individual human brain gives rise to an individual human mind.

Someone might object that we can proceed by brute force (Sandberg and Bostrom 2008): just measure the position, kind, and connections among all the neurons, and then simulate their individual behavior. Simulate all the known details, and human cognition will arise. But, even if we set aside the concerns I listed in discussing previous challenges, simulation rarely works that way. To simulate X, we need to know which aspects of the system give rise to X, so that we measure them and include them in the simulation. In the case of human cognition, we simply don't know enough about how the relevant kinds of neurons work together to produce it. We don't even fully understand the role of glial cells, which in the cortex are more numerous than neurons, and whether glial cells contribute to cognition. If we don't understand how a system gives rise to X, we don't know what aspects of the system we should measure and include in the simulation at any level of grain. Just measuring as many details as we can with the techniques we have is not enough to reconstruct how the system works if we don't know which details are important, which details are relevant to which functions, and which dynamics are generated by which details. In light of what I said above, our epistemic situation with respect to human brains is unlikely to improve to the point required for producing precise functionally equivalent simulations of individual human brains.

Fifth challenge: *idealization and simplification*. As I pointed out, the kind of brain simulation we are discussing is a hypothetical glorified version of the kind of computational modeling that computational neuroscientists engage in. Scientific models do not capture everything that is known about the system they simulate. In order to keep the mathematics tractable and the computational complexity manageable, scientific models include the most significant variables and leave the rest out. They simplify and idealize the systems they model. There is no reason to expect that idealizations and simplifications will ever be eliminated from our models of neural systems, precisely because eliminating them would make the mathematics intractable, the computations unfeasible, or both. Needless to say, idealized and simplified models behave somewhat differently from the systems they model.

Sixth challenge: *numerical approximations*. Scientists build computational simulations because the underlying mathematical models are analytically unsolvable. To compute the state of the systems being modeled, modelers employ numerical approximations. Such numerical approximations may introduce further discrepancies between the model and the system being modeled. Such discrepancies may also be present in the kind of brain simulations we are discussing.

Seventh challenge: *unpredictability of nonlinear dynamical systems*. Neural systems are nonlinear. The dynamics of nonlinear systems are very sensitive to initial conditions. Suppose that—*per impossibile*—a computational model included all the known and unknown details about an actual neural system, with no measurement errors, no idealizations and simplifications, and no numerical approximations needed. Even so, due to the sensitivity of nonlinear dynamics to initial conditions, the dynamics of an actual neural system diverges exponentially from the dynamics of any computational model that is based on a finite specification of its initial conditions. (A finite specification is all that is possible in practice.)

Eighth, *finiteness of computational resources*. The amount of details we simulate is bound by the cost of the simulation. The more details we wish to simulate, the more computing power is needed. This is where proponents of digital immortality bring up Moore's law, as if Moore's law will go on indefinitely and thus allow us to simulate an unbounded amount of neural details. Moore's law says that the number of transistors that can be fit on an integrated circuit doubles every 2 years. Roughly speaking, this means that computing power increases exponentially over time. The problem is that Moore's law is not a law of nature. Transistors cannot shrink forever. When circuits become too small, quantum noise makes digital computation unreliable. In addition, the smaller transistors become, the more expensive it becomes to shrink them further. As a result of both physical and economic limitations, Moore's law is already ending. There is a limit to how many details we can include in our simulations, and it won't go away.

In conclusion, simulating an individual human brain up to precise functional equivalence is and is likely to remain unfeasible for several reasons. There is a tradeoff between how much information we can collect about an individual brain and how destructive our methods are. Other limitations include how well we understand brain structure and dynamics, idealizations and simplifications in our models, approximations due to numerical methods, the unpredictability of nonlinear

dynamics, and resources needed to run a detailed simulation. These limitations may be somewhat ameliorated over time but not eliminated. Collectively, these limitations make it unlikely that we will ever simulate an individual human brain to a sufficient degree of approximation that warrants talk of precise functional equivalence.

6.3 Brain Replacement

Since brain simulation is unfeasible, let's see if we can do better with brain replacement. Replacing brain parts with prostheses is what neuroprosthetics does. There are several neuroprosthetic devices already on the market; for example, cochlear implants replace most of the peripheral auditory system and deliver auditory input to the cochlea. But neuroprosthetics aims at building standard devices that approximate generic brain functions to an acceptable degree—primarily peripheral functions such as delivering sensory inputs and transducing motor commands. Neuroprosthetics is not in the business of replacing individual brains down to their personality.

The kind of neuroprosthetics that would be needed for digital immortality is very different from ordinary neuroprosthetics. It involves replacing every part of an individual brain to reproduce the precise way in which that individual fulfills their cognitive functions, including their emotions and the idiosyncrasies of their mental life and character. In addition, the prostheses must be able to evolve over time, in the same way that an ordinary person evolves over time as a function of both their mental life and new experiences. In sum, the prostheses must be precisely functionally equivalent to the individual brain they replace.

Three questions arise, which parallel the questions about brain simulation:

(1*) Is it feasible to replace brain parts with prostheses that are precisely functionally equivalent to the parts they replace, until the whole brain is replaced?
(2*) Would a brain replacement be conscious?
(3*) Would the mind of the replaced brain survive through the replacement?

I will address questions (2*) and (3*) in subsequent sections. Here I focus on (1*).

In order to focus on (1*) I will assume, as in the previous section, that cognition is reducible to computation and information processing, and that either consciousness is reducible to cognition—and hence to computation and information processing—or consciousness is irrelevant to cognition. If this assumption is wrong, cognition requires consciousness and consciousness does not reduce to computation and information processing. This would rule out that brain replacement using extensions of current technology could provide what is needed for precise functional equivalence.

By contrast with brain simulation, brain replacement is consistent with the embodiment and embeddedness of the mind, so we can retain embodiment and embeddedness to whatever degree they are warranted. This is an advantage of brain replacement over brain simulation.

Despite this advantage, it is unlikely that we will ever be in a position to replace an individual brain with prostheses that are precisely functionally equivalent to the original, which is a necessary condition to make talk of digital immortality worth taking seriously. To see why, consider the following challenges.

First challenge: *replacement timing*. Depending on how many surgeries are involved, there won't be enough time. To make the prospect of preserving your mind through brain replacement intuitively plausible, sometimes proponents of digital immortality imagine replacing one neuron at a time. At the unrealistically fast rate of replacing one neuron—together with its thousands of connections—per second, replacing each of your approximately 100 billion neurons one by one would take more than 3000 years. That is a nonstarter. Replacing multiple neurons—together with all of their connections—during the same surgery would take longer than replacing one neuron at a time due to the many more connections that must be replaced. Therefore, replacing multiple neurons at a time would substitute fewer but longer surgeries for more but shorter surgeries—it would not speed up brain replacement enough to make it feasible. Replacing larger chunks of tissue while preserving individual neuronal connections during the same surgery would be a prohibitively complex operation that would take a very long time, and it would still have to be repeated for as many large chunks of tissue as there are at the relevant scale. In addition, the larger the chunks of tissue that are putatively replaced at one time, the less compelling the intuition that your mind would be preserved through brain replacement.

Second challenge: *tissue damage*. Inserting prostheses in a brain damages the neural tissue, both because of the surgery itself and because of the presence of an artificial device within biological tissue. Such damage is relatively manageable when the prostheses are few and peripheral, and when the goal is only a rough approximation of ordinary neurocognitive function. Progressively replacing a whole brain with prostheses is likely to cause enough damage to the brain to defeat the goal of building a system that is precisely functionally equivalent to the original.

Third: *epistemic limitations*. Ordinary neural prostheses perform generic brain functions to a degree that is good enough to improve functionality when patients have lesions or malfunctions. Ordinary prostheses are not quite functionally equivalent to neural tissue, let alone precisely functionally equivalent to the individual circuits of an individual brain. Designing and building neural prostheses with precise functional equivalence to the circuits of an individual's brain raises the kind of prohibitive epistemic challenges we already faced with respect to brain simulation. Specifically, it requires identifying the specific neural circuitry and dynamics of an individual brain and mimicking them in an artifact.

As we saw above, identifying the specific neural circuitry and dynamics of an individual brain involves a tradeoff between how much information we can collect about that individual brain and how destructive our methods are. Other limitations include how well we understand brain dynamics, idealizations and simplifications in our models, approximations due to numerical methods, and the unpredictability of nonlinear dynamics. Collectively, these limitations plus the amount of time and damage that brain replacement involves make it unlikely that we will ever replace

an individual human brain with prostheses that work sufficiently closely to the original to warrant talk of precise functional equivalence.

6.4 Consciousness

To continue assessing the prospects of digital immortality, let's suspend disbelief and pretend that we can either simulate or replace an individual brain up to precise functional equivalence. The next question is, will either a brain simulation or a brain replacement be conscious?

The answer depends to a large extent on whether computational functionalism holds. Recall that, in assessing the possibility of brain simulation and replacement, I assumed computational functionalism. More precisely, I assumed that consciousness reduces to cognition and cognition reduces to computation and information processing. It is now time to assess that assumption. What if consciousness does not reduce to computation and information processing?

Computational functionalism is one among many views about the nature of consciousness. To keep the discussion realistic and manageable, I will assume that consciousness either has a physical nature or is epiphenomenal. This assumption is favorable to the digital immortality project because, if consciousness is neither physical nor epiphenomenal, any form of mind uploading is going to leave consciousness out, which makes digital immortality is impossible.

I define 'physical' recursively, as follows:

> **Base clause**: human beings and other ordinary objects that we see, hear, touch, and otherwise publicly interact with are physical; our publicly observable interactions with such objects and such objects' publicly observable interactions with one another are physical; the properties in virtue of which such objects interact with one another are physical.
>
> **Recursive clause I**: any objects that compose or are composed of physical objects are physical; their interactions with one another are physical; the properties in virtue of which they interact with one another are physical.
>
> **Recursive clause II**: any objects that change the physical properties of physical objects are physical; their interactions that change the physical properties of physical objects are physical; the properties in virtue of which they change physical properties are physical; their interactions with one another are physical; the properties in virtue of which they interact with one another are physical.

An interaction is publicly observable just in case any competent observer can perform the same type of observation and obtain the same results (Piccinini 2003). The base clause establishes that we and other ordinary observable objects, their interactions, and the properties that participate in such interactions are physical. The second clause establishes that any ordinary objects' parts as well as the things they compose, their properties, and their interactions are physical. As a consequence, molecules, atoms, subatomic particles, galaxies, and the like are physical. The third clause establishes that any other things that make a physical difference to physical things, their properties, and their interactions are physical. This establishes that neutrinos, dark matter, dark energy, and the like are physical.

Given how I define 'physical', consciousness (as well as anything else) either is physical or is epiphenomenal relative to the physical universe—it has no physical effect on the physical universe. For, given how I define 'physical,' anything that has physical effects on anything physical is physical. Therefore, anything nonphysical cannot have any physical effects on anything physical; in other words, anything nonphysical is epiphenomenal relative to the physical universe. My reason for defining 'physical' in this way is that I find the notion of interaction or effect between something physical and something nonphysical unintelligible, and I'm not aware of any successful account of how something nonphysical could affect something physical (or vice versa).[3]

To assess whether a brain simulation or replacement would be conscious, we need to be clear about the relation between computation, functional properties, and the metaphysics of consciousness. There's been a remarkable degree of confusion on this topic. In order to outline the options, I will briefly sketch the ontological framework that I find most adequate and helpful in this context (see Piccinini 2020 for a more detailed account).

The universe is structured in levels of composition and realization. Pluralities of objects compose wholes larger than themselves and are composed by proper parts smaller than themselves. Properties of objects (including their relations) are realized by properties of their proper parts and realize properties of the objects they are proper parts of. Higher-level properties are aspects of their realizers; that is, a realizer is a plurality of lower-level properties of which a realized property is an aspect. Some properties are functional, which means that they are defined by specific physical effects that they have under appropriate conditions regardless of which mechanisms and qualities produce those effects. Functional properties are also known as powers or dispositions. Some properties are qualitative, which means that they are defined by their intrinsic character regardless of their effects under various circumstances.

The relation between qualities and functional properties (powers) is controversial. For present purposes, I assume that qualities and powers are connected, in that possessing qualities allows objects to have powers and having powers requires having appropriate qualities.

All higher-level properties are *variably* realizable, which means that different configurations of lower-level properties can realize the same higher-level property. For example, a mass of 1 kg can be realized by two masses of ½ kg each, three masses of 1/3 kg each, and so forth. Some higher-level properties are also *multiply* realizable, which means that they can be realized by different kinds of lower-level *mechanism*. For example, catching mice can be realized by mechanisms involving springs, glue, or electricity. Yet all mousetraps must operate on the same type of physical entity—mice. Some multiply realized properties are also *medium*

[3] For what it's worth, I also find the notion of nonphysical objects and properties unintelligible—except, perhaps, as the result of a botched analogy with physical objects and properties. I will be as charitable as I can towards those who purport to find the notion of nonphysical objects and properties intelligible.

independent, which means that the *media* on which they operate as well as the corresponding *effects* they have on such media are multiply realizable. For example, the same utterance can be transmitted via sound waves, electromagnetic waves, or messages written using an indefinite variety of materials. As long as such media are configured in the right way, they convey the same utterance. In this sense, utterance transmission is medium independent. Medium independence is stronger than multiple realizability, which is stronger than variable realizability. That is, medium independence entails multiple realizability, which entails variable realizability. Variable realizability does not entail multiple realizability, which does not entail medium independence (Fig. 6.1).

To illustrate with another example, a blender is a physical object that has both qualities and functional properties. Among its properties, a blender has a temperature and the ability to blend (soft enough) food and similar substances. Having a certain temperature is a higher-level quality. Blending food is a functional property, because it's defined in terms of a specific physical effect—turning unblended food into blended food. Being able to blend food is realized by the blender's lower-level properties, which include having components that stand in certain relations to one another. Being able to blend food is an aspect of the lower-level properties that realize it, because the lower-level realizers do much more than blend food (e.g., they are spatially arranged in a certain way, they mutually support each other, they reflect radiation, etc.). Being able to blend food is also multiply realizable, because it can be realized by different lower-level mechanisms. Being able to blend food is a medium-*dependent* property because it is defined in terms of specific physical changes, blending, to a specific type of physical medium, (soft enough) food. By contrast, transmitting a signal is a medium-*in*dependent functional property because it is defined in abstraction from any particular signaling medium. Any particular physical signal will be realized in a physical medium, but different physical media can realize the same signal type provided that they possess enough degrees of freedom organized in the right way.

As I have argued at length elsewhere (Piccinini 2015), computation and information processing, in the sense that is relevant here, are medium independent. Therefore, if consciousness were a matter of computation and information

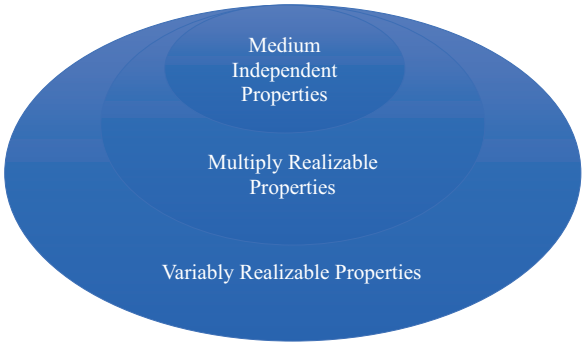


Fig. 6.1 Different Types of Higher-level Properties

processing, it would be medium independent. With this framework in place, we can define several options about the nature of consciousness:

Computational functionalism: consciousness consists solely of medium-independent functional properties.

Noncomputational functionalism: consciousness consists of functional properties that are at least in part medium-dependent.

Reductive Structuralism: consciousness consists of lower-level physical qualities.

Nonreductive Structuralism: consciousness consists of higher-level physical qualities.

Epiphenomenalist Property Dualism: consciousness is nonphysical thus has no physical effects.

There are more options than commentators usually recognize. Most of the literature focuses on the dialectic between the type-identity theory and computational functionalism, which is often conflated with functionalism simpliciter. The framework I outlined allows us to define several underappreciated options.

Computational functionalism holds that consciousness amounts to computation and information processing, which are medium-independent processes. Computational functionalism is one version of functionalism simpliciter, but there are others. Specifically, any noncomputational version of functionalism holds that consciousness reduces at least in part to (higher-level) functional properties that are medium-dependent—that is, they depend on some (possibly higher-level) physical properties.

Both versions of functionalism come in more reductive and less reductive varieties, which I lumped together in my list. Less reductive varieties of functionalism hold that the functional properties that constitute consciousness are higher-level properties; more reductive varieties of functionalism hold that the functional properties that constitute consciousness are lower-level properties.

What I dubbed structuralism is the view that consciousness is not only a matter of functional properties but also of qualities, that is, physical properties defined by their intrinsic character. Structuralism comes in more reductive and less reductive varieties depending on what level the qualities that make up consciousness are at.

The closest to standard type-identity theory is the view that consciousness is strictly a matter of lower-level physical properties, which could be functional properties, qualities, or a combination of both. Since type-identity theorists are usually vague as to which level of physical properties they identify with consciousness, I will stipulate that if consciousness is identical to properties at the neuronal level or below, then type-identity theory is correct. If, instead, consciousness is identical to properties above the neuronal level, then the type-identity theory is incorrect and some nonreductive theory holds.

This is not the right place for adjudicating between the different options about the metaphysics of consciousness. Nevertheless, we can assess what the different options entail about mind uploading.

If computational functionalism is true, it's plausible that a brain replacement would be conscious. This is because I stipulated that for a brain replacement to count as such, it must be precisely functionally equivalent to the part it replaces. Since computational functionalism holds that consciousness is a matter of

computational and information processing functions, any brain replacement that is precisely functionally equivalent to an individual brain would preserve its consciousness. As to a brain simulation, whether it would be conscious depends on whether a mere simulation (of a non-digital computation), even if it approximates the results of the original computation to a good degree, would reproduce all the functional properties that are needed for it to be conscious. Recall that a computational simulation of a brain is just a process, running on an ordinary computer, which produces successive descriptions of what a brain does. A computational simulation doesn't actually *reproduce* the computations performed by the original brain—it merely represents them. Even if computational functionalism is true, it is plausible that reproducing the original computations is necessary for consciousness. In order to reproduce them, the simulation would probably have to run on hardware that mimics neural circuitry—that is, hardware that exhibits a sufficient degree of causal isomorphism to neural circuitry. That would be closer to a brain replacement—albeit performed all at once in a location distinct from where the brain is—than an ordinary brain simulation.

If noncomputational functionalism is true, it's also plausible that a genuine brain replacement would be conscious, for the same reason as before: I stipulated that brain replacements must be precisely functionally equivalent to the parts they replace. Unfortunately, this conclusion does not support digital immortality. This is because the kind of brain replacement imagined by proponents of digital immortality is an extension of current digital technology. If noncomputational functionalism is true, consciousness is a matter of performing functions that are not computational because they are medium-dependent. We have little to no idea about what those functions might be. *A fortiori*, we have little to no idea about how to build prostheses that perform such functions. So, it follows from noncomputational functionalism that any neural prostheses capable of consciousness would require noncomputational technology that we don't have and don't know how to build. Therefore, if noncomputational functionalism is correct, we expect that any purported brain replacement that uses extensions of current digital technology would not be conscious. By the same token, if noncomputational functionalism is true, a brain simulation would not be conscious. This is because a brain simulation is just a digital representation of what a brain does. It cannot reproduce any medium-dependent functions performed by a brain.

Structuralism—whether reductive or nonreductive—has the same consequences as noncomputational functionalism. A genuine brain replacement would likely be conscious, but we have no idea how to build the right kind of brain replacement because it would have to include noncomputational technology that reproduces physical properties that we do not yet understand. Similarly, a brain simulation would not be conscious because it would not reproduce the relevant physical properties.

If epiphenomenalist property dualism is true, the follow-up question is what makes consciousness arise in the brain. It might arise due to medium-independent functional properties, medium-dependent functional properties, higher-level qualities, or lower-level qualities. Who knows? If the first of these options is correct, then

a brain replacement would likely be conscious, while a brain simulation would be conscious if it runs on neuromorphic hardware. If any of the other options are correct, a brain replacement would be conscious but we have no idea how to build one, whereas a brain simulation would not be conscious.

In conclusion, most of the options on the metaphysics of consciousness entail that a genuine brain replacement would be conscious but we have no idea how to build one, while a brain simulation would not be conscious. The two exceptions are computational functionalism and the analogous hypothesis about when consciousness arises if it's epiphenomenal.

Do we have any good reason to believe that consciousness reduces to computation and information processing? That is, do we have any reason to believe computational functionalism? I don't know of any. Most defenders of computational functionalism seem to confuse computational functionalism with functionalism simpliciter (Piccinini 2003, 2020). Therefore, I suspect that they support computational functionalism simply because they mistakenly think it follows from functionalism. In addition, the arguments for functionalism, such as the argument from the multiple realizability of cognitive functions, are much more plausible with respect to cognition than with respect to consciousness. Therefore, pending a full assessment of the metaphysical options about consciousness, I tentatively conclude that either a brain simulation or a brain replacement that uses extensions of current digital technology is unlikely to be conscious.

6.5 Survival

To complete our assessment of the prospects for digital immortality, let's suspend disbelief and pretend that we can either simulate or replace an individual brain up to precise functional equivalence, and that such a simulation or replacement would be conscious. The last question is, will the person being simulated/replaced survive through the simulation/replacement?

Whether a person survives is often discussed as a matter of personal identity: whether a future person is numerically identical to a past person. Theories of personal identity abound, and there is no consensus. Some even argue that identity does not matter for survival; what matters, instead, is that a future person be the right sort of continuation of the past person (Parfit 1984).

To avoid getting entangled in the personal identity debate, I propose to endorse a necessary condition for survival and then see if either brain simulation or brain replacement at least meets this necessary condition:

> **Survival Solitude**: A mind M survives process P only if P is such that, after M undergoes P, there can be at most one survivor.

Notice that Survival Solitude is not a sufficient condition for survival; it's merely a necessary condition. Still, if digital "immortality" violates Survival Solitude, it won't result in survival.

The point of Survival Solitude is that there is a difference between survival and duplication. Duplicating means making multiple copies of an original and being duplicated is not a way to survive, for two reasons. First, you can duplicate something while retaining the original. Second, even if duplication destroyed the original, when you duplicate something you can make multiple copies, and none of the copies has a privileged claim to being the survivor of the original in the relevant sense. If the original still exists, it is already surviving through itself. Therefore, it does not survive through the duplication process. By the same token, even if the original ceases to exist, if a process is such that it can make multiple copies of an original, the original does not survive through the process. For none of the copies have a privileged claim to being *the* survivor of the original.

To illustrate, consider photocopying a printed sheet of paper. If you can make one photocopy, you can make many. Whether you make one copy or many, none of the photocopies are survivors of the original in the relevant sense. For the original still exists! Now suppose that a photocopy machine were such that it destroyed the original in the process of making the photocopies. None of the photocopies would be survivors of the original, for the simple reason that none of the photocopies have a privileged claim to being the survivor of the original. Even if only one photocopy were made, it would still not be *the* survivor of the original, because multiple photocopies *could* have been made, and the original need not have been destroyed.

Another example is cloning an organism. If you can make one clone, you can make many. Whether you make one clone or many, none of the clones are survivors of the original in the relevant sense. For the original still exists! Now suppose that a cloning process were such that it destroyed the original organism in the process of cloning it. None of the clones would be survivors of the original, for the simple reason that none of the clones have a privileged claim to being the survivor of the original. Even if only one clone were made, it would still not be *the* survivor of the original, because multiple clones could have been made, and the original need not have been destroyed.

Someone might object that at least some entities can fission into multiple entities of the same kind. Allegedly, a single mind fissions into two minds in the case of split brains (Parfit 1984, cf. Chalmers 2010).[4] A brain includes two cerebral hemispheres, which communicate with one another primarily through a bundle of white matter called *corpus callosum*. If the corpus callosum is severed, there is compelling evidence that under some circumstances each cerebral hemisphere becomes unaware of what the other hemisphere is thinking. Under these circumstances, it is plausible that there are now two minds where there used to be only one. If the two minds that

[4] Fission thought experiments of the kind discussed by Parfit usually involve imagining that each brain hemisphere plus one half of the rest of the brain is transplanted into two different bodies. This kind of thought experiment raises roughly the same sorts of issue that I discuss in the main text for the case of ordinary split brains and has the disadvantage of being physiologically impossible for several reasons. For starters, you can't split the rest of the brain while retaining functionality in the way you can split the cortical hemispheres, and half a brain is not enough to control a whole body.

exist within a split brain count as survivors of the original, then a mind can have more than one survivor. This is a putative counterexample to Survival Solitude.

Fission is not a viable counterexample for several reasons. A small problem is that split brains are still partially integrated through the anterior commissure and other structures. In fact, under most ordinary circumstances, people with split brains behave similarly to ordinary people. Therefore, it's not clear that the two minds that exist within split brains are entirely separate from one another. A bigger problem is that neither of the partially separated cerebral hemispheres within a split brain is the survivor of the *whole* original mind. If anything, each continues "one half" of the original mind, modulo the fact that the original mind included both halves integrated with one another, whereas now each half-mind is partially separated from the other half. As a result, even if a mind were to fission into two minds through a process of brain splitting, this would not be analogous to a mind that survives as a whole. It would not be a case of survival in the relevant sense. A third problem is that this putative counterexample is only relevant insofar as it helps us think about surviving through brain simulation and replacement. But there is no analogy between brain splitting and brain simulation/replacement. Brain splitting preserves at least one half of the original brain, whereas brain simulation and replacement do not. In conclusion, fission through brain splitting is not a counterexample to Survival Solitude and is not relevantly analogous to brain simulation/replacement.

Given Survival Solitude, it is easy to see that brain simulation is a case of duplication not survival. In principle, there can be multiple computer simulations of the same entity, and the original need not be destroyed. If the original continues to exist, none of the simulations count as the survivor of the original, simply because the original is still there. Even if the original is destroyed in the process, none of the simulations has a privileged claim to being *the* survivor of the original. Whether one or multiple simulations are actually constructed makes no difference, so long as there *could* be many and the original need not have been destroyed.

By contrast, brain replacement is consistent with Survival Solitude. This is because, each time one portion of a brain is replaced by a prosthetic device, there can be only one survivor: the one and only system consisting of the original system, minus the part that gets replaced, plus the prosthesis. Since brain replacement fulfills Survival Solitude, it might be a viable option for digital immortality. Of course, this viability is predicated on brain replacement being feasible and producing consciousness using (extensions of) digital technology, neither of which is likely to be the case.

6.6 Conclusion

Brain replacement is and is likely to remain unfeasible but, if it ever became feasible, in principle it might be a way to retain consciousness and survive. Brain replacement is unfeasible for two reasons. First, replacing a small portion of a brain at a time until the whole brain is replaced would take too long and likely result in an

intolerable amount of brain damage. Second, brain replacement requires prostheses that are precisely functionally equivalent to the circuits they replace, which we have no way to build. Setting unfeasibility aside, the problem with brain replacement is that if it were based on current technology we have no reason to expect that it would preserve consciousness; in order to preserve consciousness, it would have to be based on technology that contains the physical basis of consciousness. We don't have any such technology and we have no idea how to build it, because we don't know what the physical basis of consciousness is.

Brain simulation is and is likely to remain unfeasible and it's not even a way to retain consciousness and survive. Brain simulation is unfeasible because it would have to be precisely functionally equivalent to the individual brain it simulates. But we have no way to build such a simulation. Brain simulation is unlikely to preserve consciousness for the same reason that prostheses that use extensions of current technology are unlikely to preserve consciousness. We don't understand the physical basis of consciousness, and we have no reason to expect that digital computation is a way to produce it. Finally, brain simulation is at best a way of duplicating a mind, not a way for a mind to survive.

In sum, we will not upload *our* mind to computers and, most likely, we will not upload anything *resembling* our mind to computers. When someone offers to freeze your brain so that someday they'll upload your mind to a computer, don't fall for it.

References

Blackford, R., & Broderick, D. (Eds.). (2014). *Intelligence unbound: Future of uploaded and machine minds*. Oxford: Wiley- Blackwell.

Chalmers, D. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies, 17*(9–10), 7–65.

Clowes, R. W., and Gärtner, K. (this volume). *Slow continuous mind uploading*.

Eliasmith, C., & Trujillo, O. (2014). The use and abuse of large-scale brain models. *Current Opinion in Neurobiology, 25*, 1–6.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, C., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science, 338*, 1202–1205.

Hamood, A. W., & Marder, E. (2014). Animal-to-animal variability in neuromodulation and circuit function. *Cold Spring Harbor Symposia on Quantitative Biology, 79*, 21–28.

Hauskeller, M. (2012). My brain, my mind, and I: Some philosophical assumptions of mind-uploading. *International Journal of Machine Consciousness, 4*(1), 187–200.

Hopkins, P. (2012). Why uploading will not work, or, the ghosts haunting transhumanism. *International Journal of Machine Consciousness, 4*(1), 229–243.

Izhikevich, E. M., & Edelman, G. M. (2008). Large-scale model of mammalian thalamocortical systems. *Proc Natl Acad Sci USA, 105*, 3593–3598.

Marder, E. (2012). Neuromodulation of neuronal circuits: Back to the future. *Neuron, 76*(1), 1–11.

Marder, E., Gutierrez, G. J., & Nusbaum, M. P. (2017). Complicating connectomes: Electrical coupling creates parallel pathways and degenerate circuit mechanisms. *Developmental Neurobiology, 77*(5), 597–609.

Markram, H. (2006). The blue brain project. *Nature Reviews Neuroscience, 7*, 153–160.

Markram, H., Muller, E., Ramaswamy, S., Reimann, M. W., Abdellah, M., Sanchez, C. A., Ailamaki, A., Alonso-Nanclares, L., Antille, N., Arsever, S., Kahou, G. A. A., Berger, T. K.,

Bilgili, A., Buncic, N., Chalimourda, A., Chindemi, G., Courcol, J.-D., Delalondre, F., Delattre, V., Druckmann, S., Dumusc, R., Dynes, J., Eilemann, S., Gal, E., Gevaert, M. E., Ghobril, J.-P., Gidon, A., Graham, J. W., Gupta, A., Haenel, V., Hay, E., Heinis, T., Hernando, J. B., Hines, M., Kanari, L., Keller, D., Kenyon, J., Khazen, G., Kim, Y., King, J. G., Kisvarday, Z., Kumbhar, P., Lasserre, S., Le Bé, J.-V., Magalhães, B. R. C., Merchán-Pérez, A., Meystre, J., Morrice, B. R., Muller, J., Muñoz-Céspedes, A., Muralidhar, S., Muthurasa, K., Nachbaur, D., Newton, T. H., Nolte, M., Ovcharenko, A., Palacios, J., Pastor, L., Perin, R., Ranjan, R., Riachi, I., Rodríguez, J.-R., Riquelme, J. L., Rössert, C., Sfyrakis, K., Shi, Y., Shillcock, J. C., Silberberg, G., Silva, R., Tauheed, F., Telefont, M., Toledo-Rodriguez, M., Tränkler, T., Van Geit, W., Díaz, J. V., Walker, R., Wang, Y., Zaninetta, S. M., DeFelipe, J., Hill, S. L., Segev, I., & Schürmann, F. (2015). Reconstruction and simulation of neocortical microcircuitry. *Cell, 163*(2), 456–492.

Parfit, D. A. (1984). *Reasons and persons*. Oxford: Oxford University Press.

Piccinini, G. (2003). Epistemic divergence and the publicity of scientific methods. *Studies in the History and Philosophy of Science, 34*(3), 597–612.

Piccinini, G. (2015). *Physical computation: A mechanistic account*. Oxford: Oxford University Press.

Piccinini, G. (2020). *Neurocognitive mechanisms: Explaining biological cognition*. Oxford: Oxford University Press.

Sandberg, A., and Bostrom, N. (2008). *Whole brain emulation: A roadmap*. Technical Report 2008-3. Future of Humanity Institute, Oxford University.

Schneider, S., and Corabi, J. (this volume). *Cyborg Divas and Hybrid Minds*.

Smart, P. (this volume). *Predicting me: The route to digital immortality?*

Gualtiero Piccinini is curators' distinguished professor of philosophy and associate director of the Center for Neurodynamics at the University of Missouri–St. Louis. In 2003, he received his PhD in history and philosophy of science from the University of Pittsburgh. In 2014, he received the *Herbert A. Simon Award* from the International Association for Computing and Philosophy. In 2018, he received the *K. Jon Barwise Prize* from the American Philosophical Association. In 2019, he received the *Chancellor's Award for Research and Creativity* from the University of Missouri–St. Louis. His publications include *Physical Computation: A Mechanistic Account* (Oxford University Press, 2015).

Chapter 7
Cyborg Divas and Hybrid Minds

Susan Schneider and Joseph Corabi

7.1 Introduction

Neuroscience textbooks contain dramatic cases of people who have lost their capacity to encode new memories, such as the famous case of H.M (Gazzaniga et al. 1998). Many of them suffer from severe damage to the hippocampus, a part of the brain's limbic system that is key to memory encoding. Unfortunately, these patients are unable to remember what happened to them even a few minutes ago. Theodore Berger's lab has developed an artificial hippocampus that has been successfully used in primates and which is currently in phase 3 clinical trials in humans (Schneider 2019a). If effective, the artificial hippocampus could provide individuals with the crucial ability to lay down new memories. Neural prosthetics, or "brain chips", as they are often called, are already underway for other conditions as well, such as Alzheimer's disease and post-traumatic stress disorder.

Neural prosthetics are only the beginning, however. AI-based brain enhancements are already under development. Elon Musk 's company, *Neuralink*, aims to put A.I. inside the head, merging humans and machines. Corporations such as Google, Kernal, and Facebook, as well as the US Department of Defense are also working on merging minds and machines. Neural lace, the artificial hippocampus, brain chips to treat memory and mood disorders—these are just some of the mind-altering technologies already under development (Schneider 2019a).

Elsewhere, Edwin Turner, Evan Selinger, Brett Frischmann, Cody Turner, and Susan Schneider have raised ethical considerations involving the so called

S. Schneider
Center for the Future Mind & Dorothy F. Schmidt College of Arts and Letters, Florida Atlantic University, Boca Raton, FL, USA
e-mail: sschneider@fau.edu

J. Corabi (✉)
Department of Philosophy, Saint Joseph's University, Philadelphia, PA, USA
e-mail: jcorabi@sju.edu

R. W. Clowes et al. (eds.), *The Mind-Technology Problem*, Studies in Brain and Mind 18, https://doi.org/10.1007/978-3-030-72644-7_7

mind-machine merger (Frischmann and Selinger 2018; Schneider 2019b; Turner and Schneider forthcoming). In this paper, we instead concentrate on metaphysical issues concerning *cyborg minds* – minds that are partly comprised of a non-neural substrate. Substance dualists working within various traditions have long explored the idea that the mind could transcend the brain. Now, in a different vein, proponents of the extended mind (EM) hypothesis use cases involving our interaction with technology to illustrate that the mind transcends the brain, bringing us back to the classic philosophical issue of whether the mind outruns the brain (Clark and Chalmers 1998; Clark 2003).

Brain chip cases are not the usual type of examples that are used to support the extended mind (EM) hypothesis. In their classic discussion, Andy Clark and David Chalmers generally relied on cases involving one's use of notepads and laptops—cases that remain, to this day, controversial.[1] (See Clark and Chalmers (1998).) The purpose of this chapter, however, is to urge that the brain chip cases can contribute to the EM debate in a novel way: namely, they move beyond more controversial thought experiments involving laptops and notepads and enable tests for the EM hypothesis as well as the related extended consciousness (EC) hypothesis, that is, the hypothesis that claims that consciousness is extended.[2]

This chapter will proceed in the following fashion. Section One will discuss the philosophical import of work on neural prosthetics. Then, Section Two shall offer a means of testing the EM and EC hypotheses. In Sect. 3, we turn to a more radical (and speculative) neurotechnology that is at the very early stage, and which some claim may extend the mind beyond the brain: mind uploading. This case is instructive because we've elsewhere urged that one doesn't genuinely survive uploading (Corabi and Schneider 2012; Schneider 2019). Comparing uploading to the neural prosthetics illustrates that survival (i.e., surviving as the very same person over time) and mind extension are orthogonal issues. That is, a system may not survive if it has radical changes in its underlying substrates or if the spatiotemporal evolution of the system isn't wormlike—i.e., if it does not maintain spatiotemporal continuity of an appropriate sort. How does this bear on the EM hypothesis? Intriguingly, this doesn't entail that there can't be a cognitive system that is spatiotemporally scattered and highly integrated. Such could be an extended mind. But a given biological mind may not be the sort of thing which can be radically extended and still survive.

[1] In their original paper, Clark and Chalmers do mention a hypothetical Tetris game player who had a brain implant, but it is not their main focus. In subsequent years, Clark especially began to discuss implant cases in more detail. See, for instance, Clark (2003). They still were not a focus of significant discussion until later, however.

[2] We will refer to AI-based technologies used in the brain (e.g., neural lace, brain chips, etc.) generically as 'neural prosthetics' or 'brain chips,' as other discussions of the issue tend to do, but one should bear in mind that some of these technologies may not be microchips and they may be enhancements, rather than therapies ('prosthetic' may sound therapeutic).

7.2 Transcending the Brain?

Although neural prosthetic cases were not the usual sort of examples that Clark and Chalmers originally discussed, back in Clark 2008, following up on his 2003 book, had raised a hypothetical neural prosthetic case in a published letter that was his response to Jerry Fodor's critical review of his book, *Supersizing the Mind* (Fodor 2009). The case involves a woman, named "Diva," who has brain damage and can no longer perform division. A silicon microchip is added to her brain, restoring her previous ability. As she computes, the mental process is implemented by a hybrid biological and silicon-based system. So, is Diva's mind extended then? Clark concluded: "That alone, if you accept it, establishes the key principle of *Supersizing the Mind*. It is that non-biological resources, if hooked appropriately into processes running in the human brain, can form parts of larger circuits that count as genuinely cognitive in their own right." (Clark 2008).

We agree. Indeed, this is why current work on brain chips, like the artificial hippocampus, is significant, metaphysically speaking. First, neural prosthetics are no longer hypothetical, and they are a far less controversial way in which we are cyborgs. For, to again use Berger's neural prosthetic as an illustrative example, given that the hippocampus is part of one's cognitive system, it seems implausible to deny that an artificial hippocampus, which is a device playing a similar functional role as the hippocampus, isn't part of one's cognitive system as well. As this prosthetic device becomes more sophisticated, we could see a kind of tight cognitive integration and speed of information flow that is lacking in the notebook and smartphone cases, and which resembles the way the brain interacts with its own parts.

You might object that this case doesn't extend the mind outside of the body. However, the artificial hippocampus project, to the best of our knowledge, *is* currently a project in which the electrodes link to an actual processor that is outside of the body. (The long-term goal of the research program is to locate the device entirely *inside the head*.) Further, for any sort of neural prosthetic, it seems *prima facie* plausible that if science can create an artificial part that is located in the head, then, as Clark had observed in the Diva case (2008), the device could be outside of the head and even the body, communicating wirelessly. At some point in the future, a remotely connected device could bear similar richly networked connections to the biological brain.

Work on neural prosthetics is important to the extended mind hypothesis for a second reason as well. *The extended mind hypothesis can be turned into a testable hypothesis--a hypothesis which can be confirmed by the successful use of neural prosthetics to underlie cognitive functions in the brain, or outside of the brain.* As such, it is a means of making tangible progress in the debate over EM.

Now let's consider a third, related, reason why the neural prosthetic cases are significant**.** Notice that although "mind" is a term of art, many would say that having the capacity for conscious states is a necessary condition on something's being classified as having a mind, or being minded. This indicates the centrality of consciousness to mindedness. Given this, it is interesting to ask if mind extension will be

limited to the brain's nonconscious functions, or whether neural prosthetics might replace parts of the brain responsible for consciousness. Both Chalmers and Clark are skeptical of extended consciousness, the view that consciousness extends into the world. But the work on neural prosthetics helps us better understand how, and why, consciousness could be extended. For consider that if prosthetics in areas of the brain responsible for consciousness are developed and tested, we may have indications of whether the extended consciousness hypothesis is true. For the extended consciousness hypothesis can be confirmed if neural prosthetics really replace parts of the brain responsible for consciousness without loss of consciousness.

Schneider has elsewhere developed tests for machine consciousness. One of her tests, *The Chip Test*, can be reframed as a test for the extended consciousness thesis:

> **The Chip Test for Extended Consciousness**: if neural prosthetics are substitutable in a part of the brain responsible for consciousness, then, the system has extended consciousness.

Here's the idea. Suppose that you get a silicon brain chip in part of the brain that turns out to be a substrate for consciousness. If your conscious experience ceases to function normally, this would indicate a "substitution failure" of the silicon part for the original component. Silicon just isn't the right stuff. Scientists could try and try to make a better chip. But if this kept happening, again and again, we would begin to wonder if there isn't something wrong with the silicon substitute. And if chips continued to fail, even in other substrates that were used for the development of microchips, such as carbon nanotubes and graphene, this would be a sign that neural prosthetics do not work in the areas of the brain responsible for consciousness? (Schneider 2019a).

Notice that this can occur even if neural prosthetics succeed in areas of the brain not implicated as the neural basis of consciousness (e.g., the hippocampus). In this case, the mind is extended but the EC hypothesis is false. There would be limits to the use of neural prosthetics in the brain, beyond which individuals would experience diminished or lost conscious experience.[3] And the idea that humans could merge with machines would be untenable; at best, neural prosthetics would be limited to parts of the brain not responsible for consciousness (Schneider 2019b).

On the other hand, what if the brain chips work in a part, or in parts, of the brain responsible for consciousness? In this case, we have reason to believe that the conscious mind can be extended. If consciousness can be extended through the use of neural prosthetics, this is an important result. We've noted that companies like *Kernal*, *Facebook* and *Neuralink* aim to merge humans with machines. If consciousness cannot be extended, humans cannot *fully merge* with machines, at least not in the sense the transhumanists are interested in, i.e., the sense in which one replaces their brain with AI components, either through uploading, or brain chips, so that

[3] This is not to suggest it definitively shows that sophisticated AIs we encounter would be conscious, although it is suggestive. As Schneider underscores, we can't assume, just because a silicon brain chip can allow conscious experience in humans, that AIs with the same chips are conscious, as consciousness likely depends upon the type of architectural configuration a system has.

they become forms of AI themselves. (Let's call this a "full merger" with AI.) At best, they could replace parts of the brain responsible for nonconscious activities. (Let's call this a 'limited AI integration'). Since attention and working memory plausibly involve consciousness, a failure of the conscious mind to be extended may represent a serious bandwidth limitation on cognitive enhancement, unless the brain can be radically enhanced by biological means. If this is indeed the case, then consciousness, as glorious as it is, may be the very thing that limits human intelligence augmentation, ironically. If microchips are the wrong substrate for consciousness, then A.I.s themselves wouldn't have this design ceiling on intelligence augmentation — but they would be incapable of consciousness (Schneider 2019b).

You might reply that we can still enhance the parts of the brain not responsible for consciousness. While it is correct that much of what the brain does is nonconscious computation, our working memory and attentional systems are currently regarded by many working on the neural basis of consciousness as being part of the neural basis of consciousness. Attention and working memory systems are notoriously slow, processing only about four manageable chunks of information at a time. If replacing parts of these systems with A.I. components produces a loss of consciousness, we may be stuck with our pre-existing bandwidth limitations. This could amount to a massive bottleneck on the brain's capacity to attend to and synthesize information piping in through microchips used in areas of the brain that are not responsible for consciousness. (Schneider 2019b; Turner and Schneider forthcoming).

This result would be striking: It would represent nothing less than a design ceiling on the augmentation of human intelligence: a point beyond which nonneural enhancements that attempt to replace neural tissue will inevitably fail.[4]

If the neural prosthetic cases are indeed cases of mind extension, their use turns the extended consciousness and extended mind hypotheses into testable hypotheses, hypotheses which can be confirmed. Testability is a means of making tangible progress in the debate over EM and EC, at least for cases involving neural prosthetics.

7.3 Mind Uploading

Now let us turn to the topic of mind uploading. Elsewhere, we have argued that there is little basis for thinking it likely that uploading is a way to survive, where by 'survive' we mean that the upload is really you, rather than being a being which is some sort of digital copy of you, perhaps merely preserving your memories and personality traits in a new host. (The manner in which we are using 'survival' here refers to what philosophers call 'numerical identity' rather than a looser sense of

[4] Perhaps biologically-based enhancements would be effective, however. If these can go outside of the brain or head, perhaps this would satisfy the EM hypothesis, as Katrina Vold has intriguingly suggested (Vold 2018). Indeed, perhaps they could even satisfy an extended consciousness hypothesis.

'survival' found in Derek Parfit's work, in which numerical identity does not obtain.) There are several reasons for our pessimism. First and foremost, human persons are not like computer programs or universals, being types that have different instantiations. Human persons are concrete particulars,[5] so you cannot preserve them simply by copying their abstract structure, even if that copy is causally related to the original person. There are much more stringent conditions on what is required to preserve personal identity across the sorts of transformations that are involved in uploading. In particular, we have contended that uploading does not preserve personal identity because uploading faces problems with both speed and spatiotemporal discontinuity.

In paradigmatic cases of uploading, a person's neural connections are read and then information about those connections is transported to a computer host, where the connections are then recreated or simulated in a computer environment. Even if the original brain is destroyed in the process of gathering the information (thereby eliminating an excellent rival claimant to be the true survivor and only true continuation of the original person), the transport of the information involves essentially instantaneous transport (transport at the speed of light) to a remote host and then reconstruction in that remote host. This means that the person's overall history includes spatiotemporally disjoint parts.

Consider an example. Suppose Fernando has his brain uploaded in Siena, and the computer reconstruction happens in Washington, D.C. So at time t (i.e., the time the brainscan happens), Fernando would cease to exist in Siena, at least according to those who claim one can survive uploading. Almost instantly, Fernando would shift his location and continue his existence in Washington, D.C.[6] Note that it is not the mere speed of this transition that causes problems. Unlike any normal natural macroscopic object that we know of, Fernando will travel from Siena to Washington, D.C. without traversing any of the territory in between the two. True, information about his neural connections will travel between the two, but this information will be in a highly abstract form that is not at all the same thing as Fernando. Even on the most generous understanding, Fernando will only reemerge in existence once this information has been used to create a computer copy in Washington, D.C.[7] Macroscopic objects simply do not have large spatial gaps of this sort, where they exist at one place at one time and then reemerge at a different place at a different time, without traveling in any continuous way between the two locations.[8]

---

[5] See, for instance, our arguments in Corabi and Schneider (2012) and Schneider 2019a. See also Piccinini (this volume) for related arguments.

[6] For more detail and discussion of variations, see Corabi and Schneider (2012).

[7] For Cartesians who do not believe that the seat of consciousness (the mind or soul) has a spatial location, qualifications must be added. We do believe, however, that a version of this argument will apply even to Cartesians. See Corabi and Schneider (2012). The large distance between the two sites in this example is also meant purely for illustration. If there is any distance—as surely there would be in these paradigmatic cases—the same point applies.

[8] Chalmers (2012), claims that there are objects that exist with spatio-temporal gaps of this sort. He uses the example of Yale University, which moved in 1713 from Westherfield to New Haven,

Given this, it is natural to ask: how we can maintain both that extended minds are possible, and indeed, actual, as well as deny that one could upload and survive, because a spatiotemporal discontinuity is involved? Here, it is helpful to distinguish two goals that uploading is supposed to achieve, according to its proponents: (A), uploading will be a way for an individual to survive, and hence will be a means to radically extend the human lifespan, and (B) uploading will allow for human cognitive functioning to be radically enhanced.

We see stronger grounds for optimism with respect to (B) than we do (A). If humans can be integrated with computer hosts (and sophisticated and well-integrated prosthetics become readily available), then there will likely be opportunities to enhance mental functioning beyond what is currently possible using biology alone. But this does not mean that the very same individual survives the enhancement procedure. Put another way, for all we know, it may be that there can be no extended mind which was, at an earlier point, an unextended mind. This is because no one can survive whatever enhancements are needed to truly make the mind extended. There is just too much uncertainty here, given the controversial nature of debates over personal identity in the field of metaphysics. It could turn out that the process of radical enhancement – the process of mind extension through the use of one or more brain chips—may have ended the existence of one person and created a different one (see Schneider, 2019a). Indeed, Schneider has elsewhere urged that even the addition of a single neural prosthetic could, for all we know, be too radical of a break in psychological continuity to constitute true survival. Further, the more radical an enhancement is (e.g., rewriting all of one's memories, adding new sensory modalities), the more quickly the enhancement occurs, and whether a shift in the underlying substrate (e.g., from biological to silicon), the more it seems, intuitively, that one is not the same person they were before. This is not to suggest that we can be certain that the person has ceased to exist. Rather, it is to call attention to the inherent uncertainty of the issue due to the presence of vexing epistemological and metaphysical issues that lie at the heart of debates over personal identity (Schneider 2019a). It is difficult to draw the line in a non-arbitrary fashion, locating a point beyond which we can be confident that beyond this point, one definitely does or does not survive. The extreme cases seem to be the relatively clearer ones, with the middle range cases being the most unclear. Would erasing one's childhood memories or adding new senses mean that one ceases to exist assuming one persists under normal circumstances? Uploading is an extreme case, for as we've explained, it involves radical spatiotemporal discontinuities.

Connecticut. Although some of the constituents of Yale moved from its old home to its new one, no one would claim that Yale moved continuously along the route. Hence, he contends, there is a counterexample to our claim. We believe this is wrong. The main reason is that a university is a conventional object, not a natural one (of a more "joint-carving sort") in the way a human being or human person is. Conventional objects may have strange identity and preservation conditions, and their preservation is typically very sensitive to social and contextual factors in a way that the preservation of a human being is not. See Corabi and Schneider (2014) for more details.

Further, if neural prosthetics are not able to be developed for use in areas of the brain responsible for consciousness, and these areas are key cognitive functions, like working memory or attention, it is likely that there will be a massive psychological discontinuity between the pre-uploaded and post-uploaded entities (indeed, the post-uploaded entity would not even be a conscious being). This further weakens the case for claiming that the post-uploaded entity is the same person as the pre-uploaded one. Indeed, a system lacking consciousness altogether is plausibly not even a self or person at all, as it doesn't feel like anything to be it.

This being said, the proponent of EM/EC has the following response up her sleeve. If a pre-uploaded person is spread out originally (that is, located partly inside the skull, but also partly in a smart phone, partly in a laptop, and partly in a variety of hosts around the internet), then when the parts of her that *are* located inside the skull are read, the original brain destroyed, and the neural relationships reconstructed in a computer host, there will be no impressive discontinuity. All along, after all, the person was spread out, so there is allegedly little more significance in losing the brain part of the person and reconstructing it elsewhere than there is in a favorite website used for information storage having its host server replaced one day by a new one thousands of miles from where the old one was housed. In the following section, we respond to this claim.

7.4 The Extended Mind: The Challenge

The cases we have discussed thus far—such as the example of Fernando—are paradigmatic cases of uploading. That is, they involve an original person who is a typical biological entity (made up entirely of natural building blocks of the sort that generally compose natural organisms, and if there are any non-physical elements of the person, they bear a straightforward relationship to the physical ones). This person's original brain is destroyed during the uploading process, and her brain organization is recreated in perfect detail in a remote computer host or out of remote inorganic building blocks.

But advocates of the EM hypothesis would point out that some uploads may not be like these paradigmatic examples. We don't even have to venture into science fiction territory to get a feel for some of these instances, in fact. Some philosophers have contended that normal use of current technology qualifies—in particular, people's tendency to access information on the internet using smart phones and computers, as well as store ideas and memories there. Building on arguments that were first introduced by Clark and David Chalmers,[9] they claim that a person who stores a to-do list on a smart phone app or who refers to type-written notes on Google Drive is not substantially different from someone who stores these things for recall in the brain. After all, the information is available for immediate retrieval either way

[9] See Clark and Chalmers (1998).

and can function as part of a person's practical reasoning and day-to-day decision making.[10] Further, a number of these philosophers have noted that many people are already integrating smart phones and computers into their daily processing of information, going well beyond to-do lists and notes. Some people are engaging in substantial "lifelogging"—automatically recording information about day-to-day life and storing it on a remote cloud or computer host—and this trend is growing all the time, with the proliferation of devices like Fitbits. Individuals engaged in the use of such devices, as well as more bread-and-butter forms of data storage like note-taking on cloud-based smart phone apps, are engaged in the formation of what might be called "internet extended selves" ("IESs" for short).[11] In other words, they are engaged in the process of outsourcing data storage (and to some extent data processing) to devices located outside the brain.

From here, these philosophers claim that since people are already integrating a great deal of extra-brain processing and data storage into their mental lives, much of the self already extends beyond the physical brain. They suggest that in an important sense, many of us are already uploading in 2020, and the locations of our minds already extend well beyond the edges of our skulls. But if our minds are already extending to smart phones, computers, and the internet, then, according to them, there is no dramatic shift associated with moving our minds completely to these kinds of locations, as would happen in the complete uploading scenarios that futurists envision.

Consider, then, this argument against our pessimistic view that uploading would not preserve personal identity or produce a continuation of the uploaded person[12]:

A. Many people are already forming internet extended selves (IESs).
B. IESs are split spatially between the biological brain and a variety of internet hosts (accessed typically through a smart phone or computer, but sometimes through more direct, high-tech interfaces).
C. If (A) and (B), then many people are already spatially extended in much more dramatic ways than we typically realize.
D. If many people are already spatially extended in much more dramatic ways than we typically realize, then the uploading of a person to an entirely computer

[10] To see an extreme example of such an individual, consider the case of Deacon Patrick Jones. Jones suffers from anterograde amnesia due to TBI that was a result of a large succession of concussions. In a high-tech version of the strategy in the film *Memento*, he uses a combination of high-tech software to deliver relevant information to him at all times, in spite of the fact that he cannot remember anything for more than a few seconds. Our thanks to Rob Clowes for pointing out this example. See Clowes (2013), Heersmink (2016), and https://www.psychologytoday.com/us/blog/kluge/200812/what-if-hm-had-blackberry

[11] We owe this terminology to Rob Clowes, as well as much of the inspiration for reexamining our own position in light of these ideas about the extended mind and uploading. Clowes has been at the forefront of the movement to develop arguments along these lines. See Clowes and Gärtner (this volume).

[12] This argument was inspired by a presentation given by Rob Clowes at the "Minds, Selves, and twenty-first Century Technology Conference" at the New University of Lisbon in June of 2016. We do not know if he himself would endorse the argument as stated, however.

based (or at least non-brain based) host is a less dramatic change than we have acknowledged.

E. If the uploading of a person to an entirely computer based (or at least non-brain based) host is a less dramatic change than we have acknowledged, then speed and discontinuity are not good reasons to think that the resulting upload is not numerically identical to the original person (or something close).[13]

So, by a string of MP:

F. Speed and discontinuity are not good reasons to think that the resulting upload is not numerically identical to the original person (or something close).

(A) and (B) have already been discussed at length. The motivation for (C) is that it is generally thought that people are located entirely where their bodies are—in fact, entirely where a very specific part of their bodies is: their brains.[14] However, if we acknowledge that IESs are real and are spatially split in the way that (B) acknowledges, people are then spread out in very significant ways, far beyond their brains or even their entire physical bodies. One could object, of course, that IESs are not really persons or selves. But to this point, one could reply that IESs satisfy the conditions usually set out for forming extended minds, and hence for forming real selves. These conditions, now generally called the "Trust and Glue Conditions," are:

> **Constancy**: People carry the devices they use for storage and processing with them more or less everywhere.
>
> **Direct Availability Without Difficulty**: People have routine, fast, effortless ways of accessing the information that they store and analyze with these devices.
>
> **Automatic Endorsement**: People tend to have a strong disposition to reflexively trust the information they get from their devices of choice—e.g., they will automatically trust websites with good reputations in their eyes to settle arguments.
>
> **Past Endorsement**: People have a substantial history of trusting the information they retrieve or obtain from their devices of choice.[15]

The defense of (D) and (E) is that our original examples (such as the Fernando case discussed earlier) always involve clear movement during the uploading process—the person is located entirely in one spatial region at first, and then the person (or at

[13] The "something close" qualification here is designed to include cases where a person has fissioned—as plausibly occurs in situations where an embryo splits into identical twins. The relationship the later twins bear to the original embryo would be included here as relevantly close to numerical identity. (This is what in Corabi and Schneider (2012) we referred to as "continuation.")

[14] Small qualifications are required to deal with Cartesianism, but these do not affect the substance of the points being made. The physical entities relevant to mentality—and hence, according to most Cartesians, relevant to the self—are still typically thought to be located entirely in the body or just the brain.

[15] See Clark and Chalmers (1998) for a discussion of these conditions.

least a candidate for continuation of the person) is located in an entirely different region later, without the person (or anything like a continuation or even duplicate of the person) traversing the intervening space. But if a pre-uploaded person is spread out originally (located partly inside the skull, but also partly in a smart phone, partly in a laptop, and partly in a variety of hosts around the internet), then when the parts of her that *are* located inside the skull are read, the original brain destroyed, and the neural relationships reconstructed in a computer host, there will be no impressive discontinuity. All along, after all, the person was spread out, so there is allegedly little more significance in losing the brain part of the person and reconstructing it elsewhere than there is in a favorite website used for information storage having its host server replaced one day by a new one thousands of miles from where the old one was housed.[16] So, do arguments such as this one work? Do they show that our initial pessimism was misguided after all?

7.5 Response

The argument does not license the optimism that its proponents suggest. Premise (C) is highly problematic (or perhaps (D), depending on exactly how the term 'dramatic' is interpreted). While it is true in some sense that the use of high-tech devices spreads out our cognition spatially in a way it was not many centuries ago,[17] it does not spread out our cognition in a way that is relevant to the argument for several reasons.

First, if the neural basis of consciousness—the brain processes that actually directly give rise to consciousness itself—is spread out across all these various hosts and devices, then yes, there may not be a dramatic difference between a pre-uploaded individual and a post-uploaded individual. After all, the pre-uploaded individual consisted of a biological brain interacting as the one biological node in a network of various high-tech devices, while the post-uploaded individual consists of no biological brain—just a computer host interacting as one node in a network of several high-tech devices. But, as already discussed, we do not know if extended consciousness is possible, and we have good reason to think that, even if it is possible, it is quite far off. Though brain chips of an impressive variety are already here,

[16] There may still be discontinuity right now, since the brain is still playing a major role, but the idea is that this role will decrease as time goes on as the sophistication of the devices we rely on increases. But we will not have to wait *that* long to see this decrease—nowhere near as long as the time required to produce the sorts of technologies required for uploading to take place in the way envisioned by sci-fi enthusiasts.

[17] Though this might not be true of more recent centuries, when carrying around paper notes was feasible. In the case of paper notes, the dispersal is not generally quite as impressive of course, but there is no reason why it could not be in principle and sometimes wasn't in practice, particularly once railroad, auto, and air transportation became feasible

these are exclusively targeting areas of the brain that do not plausibly play an intimate role in conscious processing.

Second, the argument moves from the presence of an EM to the claim that a given individual survives, and we've argued that this is an additional leap that requires a good deal of further discussion and argumentation.

Third, thinking of the biological brain as just one not-particularly-central node in a network of interacting devices that together comprise the self is just not a realistic way to conceive of what is going on in actual "extended mind" cases. In particular, in comparison to the high level of functional integration between components of the brain, there is very little functional integration between the devices that are outside the skull and sophisticated processing happening inside the skull. These devices, in other words, interact in very simple ways with the brain, and generally the feed forward and feedback connections are few. A smart phone heart rate recording app, for instance, typically registers one's heart rate via one input node, and then feeds the information back to the brain via one output device—generally a visual display that one can look at on a smart phone. This stands in contrast to systems within the brain—the myriad of connections involved are so complex and sophisticated that we are only now beginning to understand them in most cases. But much like peripheral areas of information registration and storage within the body, such as glucose detectors in the pancreas, the very minimal functional connections between these outside devices and the brain mean that there is not enough integration for those external devices to make a direct contribution to consciousness. Peripheral information stored in computerized devices is perhaps enhancing the richness of our cognitive lives and helping us to maintain valuable connections with our pasts and the world around us, but it is not contributing anything essential to the mind's very nature. Thus, when an upload happens, there is no reason to think it is not as dramatic and discontinuous a spatial shift as common sense would suggest.

7.6 Objections

Optimists about the extended mind's ability to produce survivors during uploading are not likely to be fully convinced by our response as stated to the preliminary considerations—i.e., the ones that involve the use of notebooks, smart phones, and slightly more advanced kinds of technology (such as lifelogging). We anticipate several objections that will arise:

> Objection A: *how do we know that the information that's stored remotely isn't entering into the sort of intimate information processing that gives rise to consciousness? If it is, what principled basis can there be for claiming that it is somehow less important to the self than brain regions involved in conscious processing?*

Response. Although they are clearly more functionally integrated than external devices, even peripheral brain regions are plausibly divorced from conscious processing, in the sense that they might feed information to consciousness but they play

no role in generating consciousness. All the more dramatically so for things outside the brain. Interestingly, as we discussed earlier, the most impressive external devices are brain implants of various sorts, such as the artificial hippocampus being developed. Although it is true that we do not know exactly what areas of the brain and processes within those areas are crucial for consciousness, the areas where engineers have made the biggest breakthroughs are plausibly many of the areas that are most clearly involved in non-conscious brain processing. In any case, there is a dilemma here. Either there is a big difference between peripheral brain regions that (e.g.) encode episodic memories and internet storage of (e.g.) videos of events in our lives, or else there isn't. If there is, the only plausible reason is that these peripheral brain regions play an intimate role in generating consciousness (i.e., they are integrated very fully in the production of conscious states), while the internet storage sites are not. If, on the other hand, there is no big difference, this is because neither the brain regions nor the internet storage sites seem directly integrated into our conscious lives.

> *Objection (B)—Regardless of how important to consciousness the peripheral information is, all that really matters to the preservation of numerical identity (or something close) is psychological continuity between the pre-uploaded person and the post-uploaded person. This will be preserved in an upload because the computer host will duplicate all the neural connections in the original brain, as well as the connections with any outside devices or neural prosthetics associated with the original. This will ensure that all memories (including autobiographical memories), personality traits, and thought patterns are preserved.*

Response. This objection leads directly into difficult issues surrounding personal identity that have been widely debated elsewhere and which we discussed earlier in the paper.[18] First, it is not even clear that uploading would preserve psychological continuity (Schneider 2019a). Certainly, it would preserve specific elements of psychological continuity, but preserving psychological continuity in a full sense may require preserving important causal connections with specific concrete entities and events. Preserving a person's memory of a child's birthday party, for instance, may require various neural states to be caused by that birthday party (and caused by that party in an appropriately intimate way). It may not be enough to have neural states (or whatever the isomorphic computer equivalent is of neural states) that merely reproduce the qualitative feel of watching scenes from the party, even if they are ultimately caused in some abstruse way by that party itself. Second, even if we grant that psychological continuity is being preserved in uploading cases, more than psychological continuity is required for personal identity to be preserved (or anything close). All that is guaranteed in an upload case is the creation of a kind of functional duplicate of the original (and, if all goes well, this duplicate will at least reproduce the qualitative mental states of the original as well). This is a far cry from preserving numerical identity, since there can be many functional duplicates of the original simultaneously and in addition the sorts of intimate causal connections we discussed just above may be missing. It was precisely the extended aspects of the

[18] See, for instance, Parfit (1984) and the expanse of literature that it inspired. We discuss these issues a bit more in Corabi and Schneider (2012) and (2014).

original person that were supposed to help provide whatever anchor was required (in addition to psychological continuity) to preserve identity. But we have seen that these external, extended aspects do not do the job the optimists were hoping for.

7.7 Conclusion

While we are optimistic about the future prospects of biomedical technologies to improve human welfare, it is far from clear that they will yield what futurists most want: a path to wholly transcend our biological embodiment. We have just seen that the above objections do not succeed. Further, it is important to bear in mind several considerations. First, it is useful to distinguish the capacity of a given neural enhancement technology to augment intelligence from its capacity to maintain personal identity over time. In principal, the former could occur without the latter. Second, we've seen that AI enhancements may allow us to test the extended mind hypothesis, and even see if the stronger hypothesis that consciousness is extended is correct, as per the aforementioned chip test. This is exciting, but at the same time, it is important to bear in mind that AI-based brain enhancements may hit a wall when it comes to the neural basis of consciousness. This wall, if it exists, would suggest that extended consciousness is not technologically feasible. Further it means that one doesn't genuinely survive if some or all parts of one's neural basis of consciousness are replaced by AI technologies. One would no longer be a conscious, minded being, and this seems essential to the survival of the mind, self or person over time. Biological enhancements of these brain functions/parts may still be feasible, however, insofar as they enhance while, at the same time, preserving the biological structures responsible for conscious experience.

References

Chalmers, D. (2012). The singularity: A reply. *Journal of Consciousness Studies, 19*.

Clark, A. (2003). *Natural born cyborgs: Minds, technologies and the future of human intelligence*. New York: Oxford University Press.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*, 10–23.

Clowes, R. W. (2013). The cognitive integration of E-memory. *Review of Philosophy and Psychology, 4*, 107–133.

Clowes, R. and Gärtner, K. (this volume). *Slow continuous mind uploading*.

Corabi, & Schneider. (2012). The metaphysics of uploading. *Journal of Consciousness Studies, 19*, 26–44.

Corabi, & Schneider. (2014). If you upload, will you survive? In R. Blackford & D. Broderick (Eds.), *Intelligence unbound: The future of uploaded and machine minds* (pp. 131–145). Hoboken: John Wiley & Sons.

Fodor, J. (2009). Where is my mind?: Review of *Supersizing the Mind*. In *London review of books*.

Frischmann, B., & Selinger, E. (2018). *Re-engineering humanity*. Cambridge: Cambridge University Press.

Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (1998). *Fundamentals of cognitive neuroscience*. New York: W. W. Norton.
Heersmink, R. (2016). Distributed selves: Personal identity and extended memory systems. *Synthese*, 1–17.
Parfit, D. (1984). *Reasons and persons*. Oxford: Oxford University Press.
Piccinini, G. (this volume). *The myth of mind uploading*.
Schneider, S. (2019a). *Artificial you: AI and the future of your mind*. Princeton: Princeton University Press.
Schneider, S. (2019b), Should you add a microchip to your brain?. NY Times.
Turner, E., & Schneider, S. (forthcoming). The ACT test for AI consciousness. In M. Liao & D. Chalmers (Eds.), *Ethics of artificial intelligence*. Oxford: Oxford University Press.
Vold, K. (2018). Overcoming deadlock: Scientific and ethical reasons to accept the extended mind thesis. *Philosophy and Society, 29*, 489–504.

Susan Schneider writes about the nature of the self and mind, especially from the vantage point of issues in philosophy, AI, cognitive science, and astrobiology. Within philosophy, she has explored the computational nature of the brain in her academic book, *The Language of Thought: a New Direction*. More recently, she defended an anti-materialist position about the fundamental nature of *mind.* In her new book, *Artificial You: AI and the Future of the Mind*, she brings these topics together in an accessible way, discussing the philosophical implications of AI, and, in particular, the enterprise of "mind design." Her work in philosophy of AI has now taken her to the Hill (Washington, DC), where she will meet with members of Congress on AI policy and organize educational events for Congress and staffers in conjunction with the Library of Congress on a range of topics, such as data privacy, algorithmic bias, technological unemployment, autonomous weapons, and more. Schneider appears frequently on television shows on stations such as *PBS* and *The History Channel* (see below for clips) as well as keynoting AI ethics conferences at places such as Harvard and Cambridge. She also writes opinion pieces for the *New York Times*, *Scientific American*, and *The Financial Times*. Her work has been widely discussed in the media (see "media" above), at venues like *The New York Times, Science, Big Think, Nautilus, Discover*, *and Smithsonian. S*he is currently working on a new book on the shape of intelligent systems (with W.W. Norton as the N. American publisher, and OUP as UK publisher.)

Joseph Corabi is a professor of philosophy at Saint Joseph's University in Philadelphia. He works on a variety of issues in philosophy of mind, epistemology, and philosophy of religion, and has published numerous articles on philosophical issues in artificial intelligence.

Chapter 8
Slow Continuous Mind Uploading

Robert W. Clowes and Klaus Gärtner

8.1 The Plausibility of Uploading

The once science fiction idea of mind uploading has long been a favorite of philosophers working on personal identity.[1] More recently, it has been floated as a serious possibility by researchers for which individual human beings and wider society needs to plan. On some interpretations the technology to do this is a nearby possibility (Bostrom and Sandberg 2008). Apart from questions of individual survival, some relate the possibility of uploading to the survival of the human race. Discussing the singularity, i.e. the hypothetical moment when artificially intelligent computers become more intelligent than humans, David Chalmers writes: "To determine whether we can play a significant role in a post-singularity world, we need to know whether human identity can survive the enhancing of our cognitive systems, perhaps through uploading onto new technology. These are life-or-death questions that may confront us in coming decades or centuries. To have any hope of answering them, we need to think clearly about the philosophical issues." (Chalmers 2010, p. 10).

The most widely entertained upload scenarios involve a three-step process. First, take a scan of someone's brain in order to record and to produce a highly detailed, indeed a functionally complete, informational encoding of the current neural state

[1] E.g. Parfit (1984).

R. W. Clowes
Instituto de Filosofia da Nova (IFILNOVA), Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa, Lisbon, Portugal
e-mail: robertclowes@fcsh.unl.pt

K. Gärtner (✉)
Departamento de História e Filosofia das Ciências; Centro de Filosofia das Ciências da Universidade de Lisboa (CFCUL), Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal
e-mail: kgartner@fc.ul.pt

R. W. Clowes et al. (eds.), *The Mind-Technology Problem*, Studies in Brain and Mind 18, https://doi.org/10.1007/978-3-030-72644-7_8

of that individual. Second, convert this into a software facsimile of the brain capable of being run as a program on a digital computer. Third, install this software - simulation, instantiation or actualization, depending on your metaphysical predispositions - onto a digital medium, i.e. on a computer, and initiate the program. Writing about such a scenario Nick Bostrum comments that: "If successful, the procedure would be a qualitative reproduction of the original mind, with memory and personality intact, onto a computer where it would now exist as software" (Bostrom 2009, p. 208). This recipe, so it has been claimed, offers the possibility – or even the promise – of individuals surviving biological death and perhaps even conducts the human race toward a form of immortality.[2]

Yet, while this gives a fair approximation of what Schneider and Corabi call "paradigmatic cases of uploading" (Schneider and Corabi this volume), or we might say vanilla mind uploading, there are other possible ways in which mind uploading may happen. This paper focuses on a form of uploading we will call slow continuous uploading or, for reasons we shall come to discuss, gradual non-destructive mind-uploading. In a nutshell, the scenario entertained here is that our brains need not be destroyed, or even scanned, in order for the uploading process to take place. Rather, as we interact with the gadgetry of the distributed internet, our minds are already undergoing a form of partial uploading which may, at some point in the future, result in the reproduction of our minds in a digital or uploaded format.

This alternative concept of what mind uploading might be is paying much attention to some contemporary technological developments alongside some contemporary theorizing around the extended mind. Especially important here is the thought that we may be Internet Extended Persons (Clowes 2020). The paper is structured in the following way in order to examine these ideas.

In Sect. 8.2, we examine some powerful objections to the standard scenario of mind-uploading. First, we examine what it might mean for a person to continue to exist through uploading; especially when we consider what a person is in terms of the traditional notion of a substance. On this metaphysically traditional account, persons are understood as enduring beings that gain their identity for instantiating certain essential properties. In an earlier paper, using this substance-based notion of person, Corabi and Schneider (2014) claim that, since uploading involves spatial and temporal discontinuties which do not reflect the way that substances normally persist, the uploading of persons violates our normal understanding of the conditions for the persistence of objects. On these grounds - and insofar as the notion of a person can be understood through the traditional notion of a substance - persons cannot be considered to survive the uploading process. More devastatingly, especially for those who hope that mind uploading might at least offer some form of continuation, Corabi and Schneider argue that this too is a vain hope. They argue that since a person neither survives nor is continued through the "uploading" process, any person voluntarily undergoing it is really committing suicide.

[2] See Bostrom and Sandberg (2008) and the article by Paul Smart in this volume.

Given the problems of such "vanilla uploading scenarios", in Sect. 8.3, we consider some less well-behaved cases of potential uploading instead. Especially, we look at the possibilities for personal continuation made possible using some current and expected near-future digital tools which may play a role in constituting us as persons. We start by considering an alternative route to personal continuation which differ from the rapid and discontinuous sort considered by Corabi and Schneider. We call this route slow continuous uploading (SCU). These scenarios depend heavily upon the idea of the Extended Mind. This idea implies that as we come to use and rely upon some tools in such pervasive ways that this can be thought of as second nature, they could come to count as proper parts of our mind (Clark and Chalmers 1998). If this view is correct, SCU might be immune to the metaphysical objections about the nature of persons staged by Corabi and Schneider, for the uploading of extended minds may be neither temporally or spatially especially discontinuous.

It might be claimed that while the idea of the Extended Mind gives us the motivation to think that cognition can extend beyond the body, it does not show that crucial aspects of the persistence of persons can actually be extended. Section 8.4, then, develops in detail some scenarios in which we examine what the relevant aspects of personhood might be and whether they can continue to persist through SCU. This section examines three ways of construing what persons are and what can be partially instantiated by digital systems that separately might count as proper parts of a person.

Section 8.5 focuses on whether slow continuous scenarios of uploading can really be mind-uploading scenarios at all. The basic objection here is that although some aspects of persons may be uploaded, the cognitive and/or conscious core is not. If this is correct, so the claim goes, such uploading would be partial at best and would not guarantee conditions of survivability or continuation. To examine this claim, we look at three possible lines of argument that cast doubt on this claim. These arguments turn on 1) questioning whether we can really sensibly draw a distinction between core and peripheral aspects of cognition and consciousness in a way that respects some contemporary ideas about the distributed nature of the human mind; 2) whether the metaphysics of consciousness can always sensibly separate core from peripheral systems; and 3) whether there may indeed be possibilities for directly uploading the core of the human mind.

Finally, in Sect. 8.6, we consider the possibility of what we call butterfly selves – i.e. selves or persons that can survive uploading even if the transition appears highly discontinuous. We find that the arguments are not so cut and dried that the question of whether an individual person may be considered to survive is resolved. We think this may depend largely on how we construe consciousness and what roles we are prepared to allow for digital system in the supervenience base of persons.

8.2 The Metaphysical Difficulties of Vanilla Uploading

In a series of papers and a recent book Susan Schneider and Joe Corabi argue that the metaphysical foundations that are assumed by many proponents of uploading are decidedly shaky (Corabi and Schneider 2012, 2014; Schneider 2019, Schneider and Corabi this volume). Their writings can be situated against a backdrop of discussions around transhumanism and the singularity in artificial intelligence that take mind uploading very seriously (Bostrom and Sandberg 2008; Chalmers 2010; Kurzweil 2000). Yet, if they are right, then there are cogent reasons to think that mind-uploading is a decidedly less attractive proposal than many proponents believe. Corabi and Schneider (2014) argue that whatever entity was created as a result of the upload process should not be considered as numerically identical to the person who consented to have their brain scanned. That is to say, uploading – at least any form of uploading that would involve the destruction of the brain – would not be *survivable*. They argue that uploading as standardly conceived is really a form of suicide.

To make their case, Corabi and Schneider introduce some metaphysical considerations about the nature of personhood, especially their numerical identity and their continuation conditions.[3] Much discussion in metaphysics construes persons as being types of objects that get their identity conditions over time by bearing properties. The metaphysical notion of *substance* is used to indicate a type of entity that is the bearer of properties in this way.[4] A person is a substance in virtue of bearing certain determinable properties.[5] Of course, any understanding of persons as substances – indeed any account of persons – must admit that persons undergo change. According to Jonathan Lowe (2009), persons or selves are "simple substances", where substances in relationship to persons means that "[…] we are persisting bearers of qualities and relations, in respect of which we continually undergo qualitative and relational change [...]" (Lowe 2009, p. 79). Consequently, persons possess certain properties, for instance they have experiences, or memories or are rational. In order to be considered the very same person, at least, some of these properties need to be maintained over time. There are however a number of different ways of construing what properties are essential to a person, and these are controversial.

[3] One of the earliest and most detailed current accounts of uploading (Bostrom and Sandberg 2008) excludes such considerations on the grounds that there are independent reasons to consider uploading in the context of whole brain emulation. Still, if the process of whole brain emulation of an individual's brain will not preserve the identity of that individual, this issue ultimately cannot be avoided.

[4] In fact there are two principal ways that we can think of substance, either as bundles or as instantiated by substrates (Schneider, 2013). Much of this discussion goes beyond the scope of this paper but we will assume that insofar as persons can be conceived of as substances it is easier to make sense of them in terms of substrates bearing properties.

[5] Corabi and Schneider are not unusual in using this framework, e.g., Jonathan Lowe (1996, 2001, 2009) develops a detailed account. A major reference point for the metaphysical debate is Descartes who bequeathed the notion of substance with his famous res cogitans.

Which properties are essential and which contingent will in part depend upon the view of the type of substance that a person is. Corabi and Schneider (2012) mention four rather different theories of what kind of substance persons are taken to be, namely soul theories, psychological continuity theories, materialist theories and also the no self-view according to which the person (or self) is a kind of fiction.[6]

The contention that mind-uploading is impossible arises from the claim that the uploaded person is neither identical to the original person, nor a continuation of it. Now, the notion of identity refers in this context simply to the question of whether the uploaded person is the "*very same person*" (Corabi and Schneider 2014, p. 132) as the original one, i.e. whether the original person survives uploading. Continuation is a bit trickier. For an entity to continue in a psychological sense, we might need to understand what is needed to be the continuation of psychological properties, such as memories and experiences, or other mental states. Yet the idea of continuation is a weaker notion than survivability, i.e. it is understood that the putative uploaded mind is no longer to be considered the very same individual (numerically identical), but rather a sort of successor being. Such a successor being might be thought to share some of the concerns, memories or experiences of its predecessor while some of these have changed.

Corabi and Schneider (2014) offer us an example for this type of continuation in the case when a human embryo (they call it "Ally") that develops into twin embryos. Even though the successor twin embryos are not identical to Ally, Corabi and Schneider observe that Ally would not generally be considered to die when it splits into twin embryos. It is rather the case that there is a special relation between them, namely *continuation*. One problem here is that although the case of Ally the embryo and her successors may be rather clear, it is less clear how we are to apply this example at the level of psychological properties. What it means for a psychological entity to be continued seems to be different in kind.

With this in mind, let us consider Schneider and Corabi's arguments against the possibility of mind-uploading. They consider two scenarios: a) instantaneous, destructive mind-uploading and b) gradual destructive mind-uploading.[7] They claim that their arguments apply to both cases. Consider first the question whether one can survive mind-uploading. In this case, Corabi and Schneider think that the main problem of uploading minds is that this involves a strange discontinuity. This means that at a certain period in time the person exists as a set of psychological or physical properties of the brain (or in the case of dualism, as a soul-like entity connected to

[6] A detailed discussion of each of these theories goes beyond the scope of this paper. However, Corabi and Schneider argue that on all versions of the substance account of person mind-uploading looks similarly unpromising. For the purpose of this paper we will focus just on the psychological view as it is undoubtedly the most promising theory from the perspective of mind uploading and also it is presupposed by the idea of the extended mind as we shall go on to discuss.

[7] We will not pursue a deeper examination of these two notions here. This is due to the fact that we believe that they are not central to Corabi and Schneider's reasoning about the non-survivability of mind-uploading. However, we think that it is important to make the reader aware that there is more than one possibility to how minds might be uploaded even in vanilla scenarios.

the body), then somehow stops existing for a brief moment, and finally comes back into existence in a somewhat distant computer. This follows from the standard uploading mechanism which consists in the retrieval of information from the brain, uploading it to a computer and the processing of this information by the computer to rebuild the informational arrangement of that brain.

For Corabi and Schneider, persons – assuming they are substances in the way we laid out – cannot be considered to survive this sort of destructive and discontinuous uploading process. The problem is that "(n)ot only does this involve an unprecedentedly rapid kind of motion for a person to follow, but this sort of motion is oddly discontinuous. For it is not as though the person moves, little by little, to the computer, so that a step-by-step spatial transition from brain to computer can be traced. Since information is being uploaded, the information has to be processed and reassembled in the computer host before anything like a functional duplicate of the original brain can be obtained." (Corabi and Schneider 2012, p. 10). The problem is that objects as we ordinarily understand them do not allow for persistence conditions that can undergo such discontinuous transitions. In a later paper they write "[t]his sort of spatial and temporal discontinuity is incompatible with standard views about the endurance conditions of ordinary objects – these intuitions are much stronger than any particular intuitions about the continued existence of a person." (Corabi and Schneider 2014, p. 136). If persons really are to be understood as object-like substances, bearing essential properties, then they should obey the rules of the normal persistence conditions of objects. Since objects do not behave in the way the mind-uploading mechanism requires, persons cannot be uploaded and survive.

But what about continuation? Can the uploaded mind at least be a continuation of the original person? Corabi and Schneider think that there is little hope that this is a genuine possibility. The reason, they claim, is that continuation is in a similar way usually thought of in physical terms. This means, whether you consider Ally splitting up into two twin embryos or cerebral transplantation, continuation is "[…] based strongly on the physical continuity between the bodies of continuations (and more specifically, brain material in the brain transplant case) and the body of their predecessor." (Corabi and Schneider 2014, p. 137). They argue that a discontinuous process as in the case of mind-uploading – which also involves the switching from a biological body to a machine – does not meet these conditions.

Thus, for Corabi and Schneider uploading is only viable if it is the sort of process that does not violate the continuity conditions of beings like us. For them, any transition should not be very rapid or very discontinuous. This said, it appears to be an open question of how rapid and discontinuous upload scenarios really are, and indeed how discontinuous uploading would have to be in order to violate the persistence conditions of persons. In any case, does this mean that there is no hope for mind-uploading? In fact, there are some other options to examine.

8.3 Slow Continuous Uploading and Internet Extended Persons

The extended mind theory (EMT) offers us an alternative route to conceive of uploading. EMT holds that under some circumstances, some tightly integrated artefacts can count as parts of minds, and even proper parts of us as selves or persons (Clark and Chalmers 1998). From this standpoint , the line of thought we will pursue here is that if minds, indeed persons, are already partially instantiated in artefacts, then there may be a route to uploading that implies far more continuity than the standard "vanilla" scenario. This EMT influenced account may initially appear exotic, but it offers a possibility for a form of mind-uploading that maintains continuities of at least some important properties taken to be of relevance to personhood in ways that are different from scenarios previously explored by Corabi and Schneider.[8] Especially the concerns about temporal and spatial discontinuities of uploaded persons, at least in the sorts of non-destructive mind uploading we will now discuss, appear to substantially lessen their force. But let us take the argument step by step.

First, let us review the argument from the EMT. EMT theory argues that if an individual uses an artefact or tool in a way that tightly parallels the way parts of his brain function, and if that tool meets some further conditions shortly to be described, then that artefact can count as part of the realization base of that individual's mind. In the now famous Otto and Inga thought experiment, we are asked to consider two people: Otto and Inga who one morning set out to visit the Museum of Modern Art in New York. However, Otto has Alzheimer's disease and finds his way to the museum making use of his trusted notebook. To achieve this aim, Otto has previously noted his intention to go to the museum along with information about how to get there. According to the original argument, the notebook – an artefact - can serve as a part realizer of Otto's dispositional beliefs just in case Otto has the right sort of relationship with the artefact. The right relationship can be articulated in terms of three conditions. First is *availability*: the notebook is a constant in Otto's life and when information resident in the notebook is relevant to his actions, he will rarely take action without consulting it. Second is *accessibility*: Otto can typically access with ease the information in his notebook. He can bring it to bear in his actions and reasoning as and where needed. Third is *trust*: on accessing information from his notebook, Otto typically endorses this information and incorporates it into his activities without further scrutiny. On this analysis, at least some of Otto's standing or dispositional beliefs are realized through the physical medium of the notebook. Clark and Chalmers claim that Otto can be said to hold beliefs about the MoMA just as Inga can, albeit Otto's beliefs are physically instantiated in his notebook. In what follows we shall grant the possibility of the EM argument without further argument and instead ask what consequences this has for our analysis of uploading?

[8] Albeit see their paper in (this volume) which treats the question of mind uploading and *cyborg minds.*

On the face of it, the paper notebook, when considered as a repository for belief, is not poised for unproblematic uploading any more than the organismic systems embodying Inga's entirely biologically realized belief system. Although we might note of course that a paper notebook is – at least relative to the current state of technology – much more readily digitized than a biological brain and with less destructive consequences, it is still clear that there are problems. Perhaps Otto could come to interact with the notebook's digitalized version in ways that strongly parallel his interaction with his paper and pen system. Nevertheless, there are strong discontinuities here. Let us assume that Otto is to begin using an electronic version of his notebook in a digital media. The location of the uploaded database housing the scans, as well as the systems that would allow ongoing interactions, are all now stored in a new physical location somewhere on a server. However, it is unlikely that the detailed physical interactions between Otto and his notebook via a pen would be precisely maintained. The affordances of digital media are, at least at fine temporal and spatial scales, just different from that afforded by paper and pen. Thus, if there is an objection to uploading brains based upon spatial and temporal discontinuities, we can see little reason why there would not be a similar objection to be made with respect to a notebook.[9]

Therefore, let us consider a related scenario whereby a more digitally adept Otto now uses a smart-phone from the beginning. Upon this he has installed a number of trusted and heavily relied upon apps which he uses to track his life and activities, and as a memory store that he can consult as and when needed. Let us thus imagine that Web-Otto's device is highly available, accessible and, as we have just noted, trusted.

Now, let us once again entertain the digital upload scenario with respect to Web-Otto. The first thing to notice is that the scanning and digitization process is no longer necessary. The apps and smartphone equipment, which could now already be considered being part of the realization base of Otto's mind, do not require being uploaded because they are already realized in digital format. Of course, this is surely a slight simplification. Rather, Web-Otto could be considered as being involved in a constant process of slow partial uploading. Crucially, non-destructive uploading for the duration of his use of his smartphone apps – or at least for the time he is relied upon them – can be considered forming part of his extended mind.

At this point, the outline of the counter-argument to Corabi and Schneider will already be clear to the astute reader: if it is the temporally and spatially discontinuous nature of mind-uploading that is the source of concerns about personal survival and/or continuation, then our current use of digital technology opens up possibilities for circumventing these objections. In the case of instantaneous destructive uploading, much hangs on the fact that it is a rapid and discontinuous event. Such

[9] The reader may object at this point that there are some obvious differences in the shear complexity of uploading information from a brain to a computer not entailed by scanning a paper and pen notebook (See, Picininni this volume for a discussion ofjust what uploading brains may entail). Still, transfer from a paper and pen system to an online mediated system would entailsome discontinuities.

uploading requires an immediate, complete and unusually fast transfer of an informational encoding from one location to the next, which still has to be processed and implemented by the computer system first, before a digitally instantiated person can even exist. In the case of gradual destructive uploading, discontinuity depends on the fact that there is a "dramatic moment" (Corabi and Schneider 2014, p. 138) where the computer processes the information of the replaced part to instantiate it. However, the theoretical possibilities of the EMT offer us a form of uploading which may be neither highly discontinuous nor disconcertingly rapid but rather continuous and slow, perhaps over years or decades. Moreover, from what we have said so far, internet-enabled mind-extension may already be viewed as forms of partial upload. The question is whether such forms of partial upload could provide us with a route to survivability or continuation.

8.4 Three Ways of Extending Persons with ICTs

Up to this point, we should note that we have only very rough criteria that might help us decide - amidst all of the change that any natural or artificial entity undergoes - which properties of a person need to be maintained in order to guarantee personal identity. Only with such an understanding in place does it seem possible to say which discontinuities are purely non-essential and consequently accidental features of that person, and which are essential properties that imply the destruction of the original entity if they are not maintained. As noted, we are here assuming EMT for our argument. The standard EMT position is generally taken to assume a fairly course-grained functionalism about minds. It is in virtue of, at least, some of Otto's dispositional beliefs being stored in his notebook that his mind is considered to be extended. EMT further implies that the properties of a person that need to be maintained in order to allow either continuance or survival are psychological properties.

Psychological theories of human personal identity all acknowledge the central problem of continuity amidst change and attempt to grapple with this problem by addressing what makes continuity viable at a psychological level. From John Locke's work onwards, a central concern has been to show how changes in memory and consciousness may nevertheless allow us to identify the same person over time. The cognitive science of memory has shown how storing memories is not just a business of storing inviolate memory-traces in discrete locations. Every act of recall changes the underlying memory trace. Moreover, it is possible that we can undergo dramatic change, such as brain injury, and still be considered to survive or be a continuation of the prior same person. It is possible that a person could lose certain capabilities such as the acuity of her vision, the flexibility of her ability to learn, or the perspicacity of her memory, and still be the same person over time despite compromises to parts of that person at either a biological or psychological level. Yet, there are presumably also changes or compromises to the agent which would mean that identity is not maintained. Changes – or discontinuities – which are so profound

that they challenge our sense of an agent's or person's identity, i.e., their personal continuation.

We have argued that the EM view opens the door to the idea that at least part of the material realization of a person may be uploaded in a continuous way. This is because, in so far as a person could be partly realized by an information communication technology (ICT), she could already be considered partially uploaded. Since part of the supervenience base of a person was already digital, there would be grounds to argue that the discontinuity was not so great. However, we have so far left rather undeveloped the idea of how an external artefactual system or tool – especially an ICT system – might contribute towards being a person and how it might contribute to the problem of personal identity. To put these points differently, we want to know when it is that a system makes a cognitive or psychological contribution such that it can also be considered part of what also constitutes that agent as a person.

In fact, there are a multitude of different manners of interpreting the ways in which personhood is constituted. But as we have already noted, we shall focus here on the psychological properties. In particular, we are interested in ways in which our artefacts and tools could make a contribution to the persons we are, and hence count as part of the realization base of us as an individual mind or person. According to Clowes (2020), there are three ways that persons might be extended by technological artefacts. These are:

- Through artefactual contributions to our narrative sense of self.
- Through artefactual contributions to our embodied skills.
- Through artefactual contributions to our abilities for agentive self-regulation.

We shall look briefly at each in turn.

8.4.1 Extensions to Narrative Self

The first – and most widely discussed way - in which an ICT system might make a contribution toward personal continuation is through systems which sustain, structure, augment or indeed replace memory; especially autobiographical memory systems. In the last several years, researchers have argued that a series of artefactual systems can play a central role in extending and augmenting autobiographical memory (Clowes 2012, 2013, 2017; Heersmink 2017, 2018), by producing, for instance, evocative objects (Turkle 2007) or actually part instantiating analogues of biological memory systems in a variety of ways. Evoking John Locke's psychological theory of personal identity, and its modern reconceptualization in terms of memory, it is widely held that autobiographical memory plays a central role in who we are as persons (Kind 2015; Schechtman 1990, 2005), and thus (especially) autobiographical memory systems are instantiated in or extended by ICTs or other artefacts. It can be argued that human personhood can be extended in the same way. Real live examples of people such as Deacon Patrick Jones seem to be existing proves of how

individuals with memory problems can use ICTs to render themselves more coherent which, otherwise, they would not be able to, if they had to rely only on their damaged organic cognitive resources (Clowes 2013; Marcus 2008). ICT can help preserve personhood.

Moreover, in the hands of some researchers, ICTs have also been used as augmentation systems. Gordon Bell's work on "Total Capture and Total Recall" has the aspiration to achieve a complete digital record of an individual's life (Bell and Gemmell 2009; Gemmell and Bell 2009).[10] Related trends in the *quantified self* (Lupton 2016; Swan 2013) movement seek to digitize detailed information about individuals through their life course with ever-expanding scope. While it is clear that at least some researchers see these trends as explicit quests into enhanced forms of self-knowledge (Bell & Gemmell, 2009), they may also be understood as new forms of autobiographical memory. Insofar as such E-Memory systems fulfill the same or similar psychological functions as autobiographical memory systems, they can be understood as providing part of the realization basis of narrative self and therefore be considered a constituent of personal identity. Aside from memory functions, there are also deep trends toward using these resources for enhanced or extended forms of self-control (Clowes 2019). This brings us to our second way in which personhood might be extended by ICT technology.

8.4.2 Extended Agency Through Self-Regulation

When, in the famous thought experiment from the original extended mind paper (Clark and Chalmers 1998), Otto used his notebook to find his way to the MOMA gallery, the notebook arguably did not operate as a form of autobiographical memory. Rather it functioned as a practical tool by which Otto organized his life, planned actions and regulated himself. Insofar as it instantiated a cognitive function for Otto, it primarily served the role of an extended self-regulation tool in a way that is reminiscent of some prominent theories of the planning agency (Bratman 2000). To put this differently, Otto's notebook did not so much extend Otto's narrative sense of self but rather his central-executive capability, i.e. Otto's ability to operate as a strong agent (Clowes 2019).

Because of the Alzheimer's disease which afflicted him, Otto was peculiarly reliant on this tool to check and regulate himself with it. But arguably the rest of us are similarly reliant on a vast variety of media, material resources, and increasingly the digital – and sometimes rather immaterial seeming – resources of the internet. Just as Otto – through the use of the material artefact of his notebook – was able to plan and regulate himself, many of the rest of us now use digital technology to do much

[10] Although Bell and Gemmel's *MyLifesBits* project can be regarded as a highly specialized research project into how much detail of a person's life can be captured through digital resources, popular self-tracking movements such as *The Quantified Self* movement show the possibilities for mass involvement in attempts to digitalize evermore aspects of individual life (Lupton 2014, 2016).

the same thing (Duus et al. 2018). Perhaps the signal technology in this new trend toward personal self-tracking and regulation devices is the wearable activity tracker (or WAT) of which the Fitbit is one of the best-known brands. Yet the use of Fitbits and other WATs is just the vanguard of a host of other physical and software devices and apps which are principally designed to allow users to track and self-regulate an ever-increasing range of personal attributes and activities (Lupton 2014). Fitbits and a variety of apps, and app ecologies, that we interact through mobile devices are increasingly turning into a new sort of cognitive ecology (Smart et al. 2017), a large part of which is constituting a new regime of self-regulation (Clowes 2019).

We can frame much of what is happening through the use of this new regime of self-regulation technology around Michael Bratman's ideas on strong agency (Bratman 2000). According to Bratman, the faculties of extended planfulness, reflectiveness and self-regulation all interrelate and hang together in a sort of cognitive suite which characterizes the – as far as we know – uniquely temporally-extended form of agency we find in human beings. Such self-regulative agency is deeply bound up with our coherence and continuity as persons over time, or in other words, personal identity. Such capacities have been deeply related to personhood as such through the idea that it is only by being able to take a second-order attitude toward our believes and desires that we are able to operate as "selves" or persons in a strong sense (Frankfurt 1971, 1987). Insofar as I know myself through temporally extended activities, many of the processes of self-regulation and reflection are increasingly instantiated by extended tools. Consequently, if Frankfurt's idea of the hierarchical self is enabled by ICTs, it seems likely that core aspects of personhood can be extended. A central part of the case for SCU is that, inasmuch as we rely on ICTs to implement many of these self-structuring systems, we are – once again – already partially uploaded into the digital realm (Clowes 2020).

8.4.3 Being Someone Through Embodied Skills

Finally, a third form of personhood may be considered to depend on our embodied or situated skills. The idea here is that, as ICT systems and especially the internet become increasingly transparent resources evermore deeply incorporated in our skillful practice, we are likely to rely on them to ever greater degree. Clowes (2020) discusses some of the literature on how the identity of human persons is deeply dependent upon our embodied skills. Human skilled practices have long been understood to entail a form of deep artefactual dependence (Dreyfus and Dreyfus 1980; Heidegger 1927; Luria and Vygotsky 1992; Malafouris 2013). A potter's skills are tightly constrained by his crafty, implicit knowledge and abilities as exercised at the wheel.[11] A violinist is most herself when playing her violin. A writer

[11] See Malafouris (2008a) for an elaborated discussion of how cognition and agency are constrained, and made possible, through what he calls *material engagement.*

depends heavily on the resources of Wikipedia to structure his thoughts. An architect's imagination is facilitated, and her designs and their elaboration made possible, by the deep way she has integrated a particular computer-aided design (or CAD) interface into her thinking and creative practices. It is possible to imagine – in the latter case – that if the architect loses access to her tools, she becomes strongly compromised in the process, and thus her abilities to *be* herself. Human skilled practice is undoubtedly becoming ever more involved with deeply incorporated digital tools (Clowes 2015b; McCullough 1996). Insofar as such resources can be considered extended, and thus proper parts of our minds, then our embodied skills can be part instantiated and embodied in our tools - including ICTs – and they may also be involved in a process of slow uploading.

8.4.4 Distributed Selves and the Cloud

It is important to stress here how in each of the three ways described above, the technologies are – potentially at least – playing an active rather than a passive role. The AI-infused technologies of the contemporary internet are already highly active, both constraining and enabling a host of our cognitive abilities. Increasingly our minds are regulated by a data profile held by companies such as Google and Facebook. The "digital shadow" (Lupton 2016) that we create through the – apparently – innocent act of searching on the web leaves data about our activities with multiple digital systems. This harvested data increasingly shapes and constrains our activities. It is important to realize that this is not merely a property of our online lives, but progressively digital constraints and affordances affect all aspects of our lives. Luciano Floridi has coined the term *onlife* to convey the way that the activities we perform through digital media technologies can no longer be best conceptualized as a separate digital realm, but must be seen as an inseparable dimension of the human lifeworld (Floridi 2014, 2015).

The technologies of the mobile and pervasive internet – something one of us has called cloud technologies (Clowes 2015b) – are only deepening this trend. Digital technologies that measure the minutiae of our lives and use them to shape our activities and behavior through "nudges" (Leonard 2008) have become ever-present accompaniments to our activities. Personalized data-shadows mediated by our smart phones and an ever-expanding range of wearable gadgetry increasingly shape an expanding range of our cognitive activities and with them our lives. Much of this has happened – at least from the individual user's perspective – in a largely unconscious way. Nevertheless, this technology should not simplistically be understood as something that merely controls us, or simply constrains human agency. Cloud technology can, and indeed already has been, appropriated by many early adopters as a new type of instrument of self-regulation system, and thus potentially a means of developing plans, regulating ourselves with those plans and ultimately for self-shaping (Clowes 2019). Moreover, human beings have a long history of using technology to regulate themselves. Although the new "smart technologies" may be

altering human agency in important respects, there is a deep continuity in the history of the human use of cognitive technology, since at least paleolithic times to regulate and structure oneself (Luria and Vygotsky 1992; Malafouris 2008a).

This type of incorporation of technology arguably extends a long-term trajectory of the human species by which our minds, selves and who we are as persons do not just depend on our ostensible nature as biological organisms, but as extended beings, part realized by our tools and technology (Clark 2003; Clowes 2012; Malafouris 2008b). If the extended mind view is correct, there are reasons to think that at least certain internet apps should count not just as cognitive extensions – or even extensions of our minds – but as part of our personal substrates, i.e. part of what makes any individual that very individual.

The case that we have mounted so far is that at least some of us are already hybrid beings with hybrid minds (Menary 2007). Insofar as internet technology is playing a constitutive role in psychological properties of who we are as persons, we can already be considered *partially uploaded* to the internet. To clarify – as the notion of partial uploading may appear to be controversial – all that is intended is that if a digital resource plays a constitutive role in who a particular person is, and that digital constituant is instantiated as a resource, program or data source on the internet, then that person can be considered partially uploaded. A person with this sort of hybrid realization base is not fully uploaded because she still has biological components.

Our deepening reliance on internet-based cognitive technologies is of great moment here, since it could be argued that a substantial part of the minds of individuals has already been "uploaded" to the internet. Uploading here is not to be understood as a discreet process of capturing the brain-state or neural-encoding of an individual to be transferred into a digital medium, but rather the temporally extended process of ongoing self-regulation by a host of digital tools many now becoming the constant accompaniment of our cognitive processes. The SCU scenario can now be specified with more precision. Persons of the twenty-first century are increasingly becoming complex beings with part digital substrates. Our digital components play complex and somewhat autonomous roles in regulating our organic components such that they are becoming integral components of many human persons. It is no doubt now clear how this dialectic confronts the original arguments staged by Corabi and Schneider (2012, 2014). Insofar as the main problem for digital uploaded beings is that the process is radically discontinuous this argument may only apply to traditional "vanilla" uploading scenarios. But in the more exotic world of partial and regulative uploads we arguably already inhabit, the discontinuity worry may be significantly downgraded. Uploading can be a partial ongoing process lasting years or decades, or a lifetime. A transition to the digital life may even come to be seen as an extended and natural transition into the digital realm. Remaining discontinuities on this picture may come to seem relative, or perhaps even trivial, rather than absolute.

8.5 On Core Consciousness and Personal Continuity

In an article commissioned and written for this book, Schneider and Corabi have however thrown significant doubt over the sanguine account of SCU just offered. Specifically, they doubt whether uploading internet extended persons (IEPs)[12] are any more likely to survive uploading or even be continued by the process. This has not so much to do with the problems that some form of *partial uploading* of IEPs may be possible, but rather that such a process cannot accommodate the whole person. They worry moreover that even if some form of digital entity can be created that continues important aspects of a person, the core of that person – that person's consciousness – cannot be so-uploaded. And if this is the case, then even a form of SCU along the lines we have been examining in this article, is not a route to personal survival of biological death – or a form of continuation – but rather a form of suicide.

Schneider and Corabi do not dispute in principle the idea that the digital parts of a person may be uploaded – or indeed already are uploaded.[13] Persons, specifically IEP scenarios as we have just described, may indeed be part constituted by an extended supervenience base which extends beyond the standardly considered organismic boundaries. Rather they argue that essential constituents of a person are not thereby uploaded. The problem is that even if many aspects of the mind could be uploaded in a continuous manner as we have outlined above, the core aspects – the very aspects of a person that make that person a person – cannot. For Schneider and Corabi the problem is that the core of consciousness, the biological basis of conscious mental life cannot not be uploaded for it supervenes locally on the brain.

Perhaps surprisingly, such a view might be seen as gaining support from Andy Clark in his book *Supersizing the Mind* where he argues that while the mind is extended, consciousness is likely not (Clark 2008, p. xiv). The problem then is clear. There may be IEPs and parts of those may be realized by a digital substrate, but the very same problems remain for SCU cases as in vanilla scenarios. The problem is that the person preserving the core of consciousness cannot be uploaded. This is because the digitally instantiated part of an IEP may be an important part of the overall instantiation base of a person's extended mind, but not of that person's extended consciousness. The crucial identity conferring remnant remains part of the biological substrate, and any attempt to upload this will thus fall victim to the very same objections targeted at classical uploading accounts.

It must be acknowledged here that a great many theorists, some of whom hold radically different accounts of the self, would concur with Schneider and Corabi's

[12] This was the subject of an extended discussion from the conference *Mind, Self and twenty-first Century Technology* which was the taking place on the 22nd of June 2016 in Lisbon, Portugal.

[13] In the account developed here, parts of the person come to be instantiated in a digital system through a long process of interaction. Uploading is in some respects a rather loose term for the process as a person constituting parts of the system may start of as an external system which are integrated through interaction.

assessment of prospects that even SCU cannot guarantee the continuation of a person, let alone the survival of human beings. The central move in any such argument would be to claim that at the core of human consciousness is a biologically created and sustained sense of self (Damasio 2000; Seth 2013; Zahavi 2005). Refuting such objections might be considered *taking the hard road* with respect to claims about a core self or consciousness and yet a number of important theorists of the self and consciousness can be interpreted as pursuing such a route.

A first response develops from what our account has already implied about the distributed nature of self, namely that the idea that there are core inseparable internal components that can be drawn apart from other parts of the system is a misapprehension about the human mind. Rather, the self might be considered an emergent phenomenon that depends upon the interaction of systems both biologically internal, cultural and even – in ways that we have discussed in Sect. 8.4 – be part technologically constituted. To put this in stark terms, the self is not, on this analysis, to be found in core systems, but emerges from a set of narrative and culturally involved processes that do not have a core in any straightforward way.[14] Extended mind theory and specifically Andy Clark's idea that we are, or could be, extended selves (Clark 2006), gives us one set of reasons for denying the core-self view.

Daniel Dennett's (1991) book *Consciousness Explained*, develops a number of themes that could lend support to the idea that there is no special core to consciousness or self. A central theme of the book is that not only are there no special places in the brain – a Cartesian Theatre as Dennett calls it - where consciousness comes together, but the very suggestion is incoherent. Throughout Consciousness Explained Dennett offers a variety of reasons to think that there is not a privileged set of central systems that instantiate consciousness. He rather claims that consciousness is produced by a heterogeneous bunch of systems especially tied up with the creation of language, the ability to report (or confabulate) the existence of inner states, which give rise to, what Dennett calls, the *center of narrative gravity.* It is wrong to think there is a core to consciousness where really there are only a heterogeneous set of systems which generate multiform appearances of conscious mental life. On Dennett's account, it is because we create, more or less coherent, explanatory narratives about ourselves and our life histories that we can be said to have a consciousness self at all. For Dennett, there are no special systems that produce an inner light show of the phenomena itself, and our personal coherence over time is something of an illusion.[15]

[14] It is important to note here that there are a variety of positions with respect to the constituent of self in which narrative accounts of self and accounts based on subjective or biological cores can be seen as poles (Gallagher 2000). It is far beyond the scope of this paper to discuss the different possibilities for articulating, e.g. narrative conceptions of self or core conceptions of self as these are multiple and highly differentiated. It is also worth noting that many attempts to give an account of self uses elements of both narrative and core-self theories. See e.g. Clowes and Gärtner 2020, Gärtner and Clowes 2020, and Schechtman 2011 for discussion.

[15] When performing what Dennett calls heterophenomenology we can glimpse behind the curtain at the workings of the systems that spin consciousness (Dennett 2003, 2004).

This Dennettian train of thought can be used to argue that there are not strict divisions to be made between consciousness involving parts of the brain, and non-conscious parts of the brain. Consciousness rather is a feature of whole agents and their system of self-interpretation. Narrative capabilities are of great importance here. Indeed, on some of the accounts of IEPs above - such as the narrative account - consciousness could be closely tied to one type of system that has arguably already been part extended. Arguably, this option requires us to take a Dennettian and thus - for some realists - slightly irrealist view about consciousness. On such views, many of the systems which are involved in consciousness are precisely those that spin narratives, and as we have seen, these have been argued to be exactly the sort of systems that can be extended by ICTs (Clowes 2013, 2017; Heersmink 2017, 2018).[16]

For somewhat different reasons, Thomas Metzinger has argued that the conscious self is another sort of illusion, this time produced by a variety of control systems in the brain bent on the task of self modelling (Metzinger, 2004). On Metzinger's view, the self is virtual (Clowes 2015a; Metzinger 2009). If Metzinger is right that the self is a special sort of illusion generated by our brain's representational media, these need not play any special role in generating personal identity. From there it is difficult to see why transitioning to a digital substrate would cause special problems.[17] On Metzinger's view the idea of a person as a sort of inner substance is an illusion. Following this version of illusionism, it is difficult to see why the underlying substrate that is generating "the illusion of a conscious self" might not be part dismantled and then instantiated on another substrate. Given Metzinger's view that the conscious self is a sort of illusion generated by the cognitive system, the idea that there would be discontinues in the generation of this illusion would not be especially problematic. Indeed, part of Metzinger's argument, like Dennett's, for the illusory nature of the self is that there are very often deep but unnoted discontinuities in our sense of self. (An interesting test for degree in which both Dennett and Metzinger really hold their views is to enquire into how worried they might be by personally undergoing the sorts of uploading scenarios considered here).

But how might the "illusion of conscious self" be instantiated in a machine. For a detailed path to such a transition the reader is directed to an attempt developed as just such an account in this volume. In his paper *Predicting Me: The Route to Digital Immortality*, Smart investigates a number of models by which deep learning or predictive processing systems could develop a detailed predictive model of a human being over his or her lifetime. Smart offers different scenarios in which such a

[16] See Keith Frankish's paper in this volume for the detailed development of the idea that it is Type 2, essentially recent and specifically human reasoning systems, that give rise to the appearance of consciousness. Frankish defends the claim that these systems of the "virtual mind" are the ones that arise in cultural practices and through – following Dennett – processes of autostimulation. Like narrative systems, systems for autostimulation might be precisely the sorts of system that could be coupled with or instantiated in cloud technology. If this line of thought is correct, then Type 2 systems are precisely the sorts of systems we could expect to be uploaded through long interactions with online digital systems.

[17] See also Paul Smart's paper in this volume on deep learning, predictive processing and some ways that self-modelling might be generated.

model could be built. The deep-learning approach offers a way forward here in that it may be that a silica system can produce such a good facsimile of the organic component that there is no possible scenario in which we can discriminate any difference between it and a digital simulation. Rather than going further here into the detailed questions of whether such modelling requires some degree of fidelity, in the following section we will look at the deeper metaphysical implications of such modelling. Any view on uploading, even built upon the ideas of lifetime modelling or SCU will have to face the fact that there will undoubtedly be important discontinuities in any upload scenario. The important question is: What are we to make of these discontinuities?

8.6 On Psychological Continuity and Butterfly Selves

How much can an entity change and still be the same? How much can a person change and still be considered the same person? These are old problems in philosophy but have been cast into a new light by the scope and depth of our interactions with a range of smart technologies (Clowes 2013, 2019). We have argued in the last chapter that the new technologies of mind-extension offer a route to at least a form of personal continuation that doesn't obviously fall foul of the objections from Schneider and Corabi. This possibility largely turns on there being deep continuities between the way that we currently integrate extended digital resources into our minds and the idea of mind uploading.

But biological death – for the time being at least – is still unfortunately inevitable. Any continued existence of a mind, person or self which can be considered to cross this boundary will certainly have to be considered rather discontinuous. This raises the question of what we are to make of the continued "life" of the extended component of a person, once the central biological component has ceased to be. First, we should note that the digital part of the substrate cannot continue in a way that merely continues the hybrid person it once extended. Where the digital component merely extended the biological core it now needs to function as an independent entity.

At least one major current metaphysical position on continuation of personal identity, namely animalism (Olson 1999), must maintain that the end of animal instantiation of life is the end of the person. On the animalist view uploading of all types seems to be ruled out. This is one reason we have focused on psychological properties here. So, if continuation or survival are to happen through SCU, they will need happen on psychological grounds. Most Contemporary versions of psychological continuation theories turn, as we have noted, on memory or consciousness. These two grounds ramify rather differently. On the grounds of memory, there is good reason to think that digital systems are already playing important roles in structuring and retaining memories for at least some individuals (Clowes 2013, 2017; Heersmink 2017). It may even be that the use of memory extending

technologies mean that the mnemonic basis for self is extended beyond the normal range for IEPs (Clowes 2012).

Consciousness looks like it might be problematic. As we have seen on standard brain-based views of consciousness, it looks like there is at least a radical discontinuity between biological and uploaded selves. There are two further possibilities here. On the Frankish view of consciousness being associated with Type 2 —slow, evolutionary recent and culturally generated— systems there is a case to be made that such systems are exactly those that could be uploaded (See Frankish this volume). An alternative view discussed in detail in Schneider and Corabi is that the ultimate destination of uploading, even of the slow and continuous variety discussed here is the cessation of consciousness. If this is true, if the digital instantiation of the mind is not conscious, this would seem to be the end of continuation. But even here it is worth considering possible scenarios.

Thus, we must consider the basic question again. Given that life, memory and consciousness are in many respects replete with multiple discontinuities, what sort of discontinuities are so profound that they can be considered to block both survivability and continuation. Consider how some natural entities undergo profound change and yet are still considered the same. Consider the life cycle of a butterfly. A butterfly goes through several life stages from caterpillar, to chrysalis, to butterfly. Through the process of transformation, the caterpillar's body is broken down into a chemical soup within the chrysalis which is then the raw material for the butterfly. The change is radical in each step of transformation, and yet we do not say that the butterfly is a different individual from the egg, caterpillar or chrysalis which were its earlier incarnations. This seems to demonstrate that at least some individuals can pass through radical transformations without provoking the sense that they are not the same individuals anymore.

At this point we would like to introduce Theseia, a fictional 35-year-old human being living in the early twenty-first Century and who relies heavily upon a variety of ICT resources to the point that they might be considered part-constitutive of who she is as a person. Theseia, like other humans of all epochs, is undergoing a constant process of change, development and renewal at a variety of different temporal and spatial scales. At one level, Theseia can be understood as a biological being – an organism, or biological human. Over a ten-year period, every cell in Theseia's body is replaced by a new one. Individual cells die and are replaced, but this happens as part of a process of the self-maintenance of Theseia as a bodily organism. Arguably these changes guarantee the survival of Theseia as a person, and from this we can conclude that the changes in some parts of an entity do not undermine its continued existence.

Theseia has been a life-long user of person extending digital systems. As she now lives through the advanced stages of old age she has come to rely more and more on those systems. It is Christmas time and Theseia is reminded to send Christmas messages to friends and family members by "agent" software that now fulfils many of her commitments. Theseia relies heavily on the extended systems but they come to play new roles as she has aged and increasingly act, with her blessing and in accordance with her policies, rather autonomously. The systems now

support many of her fading cognitive abilities and support continued connections with her previous life. Now, consider that Theseia, or at least her biological parts, after a long and fruitful life, passes over. However, her digital components which she has relied upon and carefully nurtured over much of her biological life continue, now instantiated by latest AI-based personal instantiation technology.

In Theseia's case, any claims for continuation or survival between her as hybrid (exo-self extended) person and purely exo-self remainder will be highly controversial, at least for now. The discontinuities certainly at a material level appear dramatic and by the standards of the pre-twenty-first century regular human life-cycle would appear to involve a radical discontinuity, i.e. the radical discontinuity of the death of Theseia's biological body. Does this mean that persons cannot be considered to endure through such radical discontinuity? It is worth noting here that many religions across the world have considered it to be possible that some form of continuation indeed survivability was possible after the death of the biological body. The whole idea of an (possibly immaterial) spirit rested on the notion that some essence of a person could survive biological death which by necessity involved a transformation in the nature – in the incarnation – of that being. Given the conceivability of the survival of a person through a transformation of this type, can we therefore rule out all forms of mind uploading on primarily conceptual grounds?

A supporter of upload survivability on psychological grounds could hold that machine uploading of the type we examined in the case of Theseia, while not being the same kind of creature as its earlier incarnation, nevertheless should count as a psychological successor. This is because the new – digital / informational – entity has the right kind of psychological continuity in terms of memories, concerns and perhaps – at least on some models – consciousness. Theseia moreover followed a reasonable trajectory of development from biological Theseia, through cyber-Theseia – part instantiated in a series of internet-based systems deeply involved with her biological systems – to the final fully digital uploaded-Theseia.

If we are prepared to concede that at least some entities continue in some respects through major discontinuities, then the deep question here is whether persons are the sorts of entities that can survive *these* discontinuities. This may come ultimately down to a question of intuitions of what we think counts in the question of psychological continuity. If we take the Dennettian / Clarkian account of distributed and somewhat illusory sense of self seriously then it may be that, at least in slow uploading scenarios, the discontinuities that undoubtedly exist, might not be the ones that matter for a form of personal continuation. We think that, at least, the metaphysical case against SCU that we have mainly discussed in this paper has not yet ruled out the persistence of persons on these grounds.

Acknowledgements Robert W. Clowes's work is endorsed by the financial support of FCT, 'Fundação para a Ciência e a Tecnologia, I.P.' under the Stimulus of Scientific Employment (DL 57/2016/CP1453/CT0021) and personal grant (SFRH/BPD/70440/2010).

Klaus Gärtner's work is endorsed by the financial support of FCT, 'Fundação para a Ciência e a Tecnologia, I.P.' under the Stimulus of Scientific Employment (DL 57/2016/CP1479/CT0081) and by the Centro de Filosofia das Ciências da Universidade de Lisboa (UIDB/00678/2020).

This work is endorsed by FCT project "Emergence in the Natural Sciences: Towards a New Paradigm" (PTDC/FER-HFC/30665/2017).

References

Bell, C., & Gemmell, J. (2009). *Total recall: how the E-memory revolution will change everything*. New York: Dutton.

Bostrom, N. (2009). The future of humanity *New waves in philosophy of technology* (pp. 186–215): Springer.

Bostrom, N., & Sandberg, A. (2008). Whole brain emulation: A roadmap. *Lanc Univ Accessed January, 21*, 2015.

Bratman, M. (2000). Reflection, planning, and temporally extended agency. *The Philosophical Review, 109*(1), 35–61.

Chalmers, D. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies, 17*(9–10), 7–65.

Clark, A. (2003). *Natural born cyborgs: Minds, technologies and the future of human intelligence*. New York: Oxford University Press.

Clark, A. (2006). Soft selves and ecological control. In D. Spurrett, D. Ross, H. Kincaid, & L. Stephens (Eds.), *Distributed cognition and the will*. Cambridge, MA: MIT Press.

Clark, A. (2008). *Supersizing the mind*. Oxford, UK: Oxford University Press.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*, 10–23.

Clowes, R. W. (2012). Hybrid Memory, Cognitive Technology and Self. In Y. Erdin & M. Bishop (Eds.), *Proceedings of AISB/IACAP World Congress 2012*.

Clowes, R. W. (2013). The cognitive integration of E-memory. *Review of Philosophy and Psychology, 4*, 107–133.

Clowes, R. W. (2015a). The reality of the virtual self as interface to the social world. In J. Fonseca & J. Gonçalves (Eds.), *Philosophical perspectives on self* (pp. 221–276). Lisbon: Peter Lang.

Clowes, R. W. (2015b). Thinking in the cloud: The cognitive incorporation of cloud-based technology. *Philosophy and Technology, 28*(2), 261–296.

Clowes, R. W. (2017). Extended memory. In S. Bernecker & K. Michaelian (Eds.), *Routledge handbook on the philosophy of memory* (pp. 243–255). Abingdon, Oxford: Routledge.

Clowes, R. W. (2019). Immaterial engagement: Human agency and the cognitive ecology of the internet. *Phenomenology and the Cognitive Sciences, 18*(1), 259–279. https://doi.org/10.1007/s11097-018-9560-4.

Clowes, R. W. (2020). The internet extended person: Exoself or doppleganger? Limité. *Límité. Revista Interdisciplinaria de Filosofía y Psicología, 15*(22).

Clowes, R. W., & Gärtner, K. (2020). The pre-reflective situational self. *Topoi, 39*, 623–637. https://doi.org/10.1007/s11245-018-9598-5

Corabi, J., & Schneider, S. (2012). Metaphysics of uploading. *Journal of Consciousness Studies, 19*(7–8), 26–44.

Corabi, J., & Schneider, S. (2014). *If you upload, will you survive?* Intelligence Unbound*:* Future of Uploaded and Machine Minds, The, 131–145.

Damasio, A. R. (2000). *The feeling of what happens: body, emotion and the making of consciousness*. London: Vintage.

Dennett, D. C. (1991). *Consciousness explained*. Harmondsworth: Penguin Books.

Dennett, D. C. (2003). Who's on first? Heterophenomenology explained. *Journal of Consciousness Studies, 10*, 19–30.

Dennett, D. C. (2004). *A third-person approach to consciousness sweet dreams: Philosophical obstacles to a science of consciousness*. Cambridge MA: Bradford Books, MIT Press.

Dreyfus, H., & Dreyfus, S. (1980). *A five–stage model of the mental activities involved in directed skill acquisition*. Berkely, CA: Operations ResearchCenter, University of California.
Duus, R., Cooray, M., & Page, N. C. (2018). Exploring human-tech hybridity at the intersection of extended cognition and distributed agency: A focus on self-tracking devices. *Frontiers in Psychology, 9*(1432). https://doi.org/10.3389/fpsyg.2018.01432.
Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality*. Oxford: OUP.
Floridi, L. (2015). *The Onlife Manifesto: Being Human in a Hyperconnected Era*. Cham/Heidelberg/New York/Dordrecht/London: Springer.
Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy, 68*(1), 5–20.
Frankfurt, H. G. (1987). *Identification and wholeheartedness*.
Frankish, K. (this volume). *Technology and the human minds*.
Gallagher, S. (2000). Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences, 4*(1), 14–21.
Gärtner, K., & Clowes, R. W. (2020): Predictive processing and metaphysical views of the self. In Mendonça, D., Curado, M. & Gouveia, S. S. (Eds.), *The science and philosophy of predictive processing*. London: Bloomsbury. https://doi.org/10.5040/9781350099784.ch-004
Gemmell, J., & Bell, G. (2009). The E-memory revolution. *Library Journal, 134*(15), 20–23.
Heersmink, R. (2017). Distributed selves: Personal identity and extended memory systems. *Synthese*, 194(8), 3135-3151.
Heersmink, R. (2018). The narrative self, distributed memory, and evocative objects. *Philosophical Studies,175*(8), 1829–1849.
Heidegger, M. (1927). *Sein und Zeit*. Tübingen: Niemeyer.
Kind, A. (2015). *Persons and personal identity*. Cambridge: Polity Press.
Kurzweil, R. (2000). *The age of spiritual machines: When computers exceed human intelligence*. New York: Penguin.
Leonard, T. C. (2008). Richard H. Thaler, Cass R. Sunstein, nudge: Improving decisions about health, wealth, and happiness. *Constitutional Political Economy, 19*(4), 356–360.
Lowe, E. J. (1996). *Subjects of experience*. Cambridge: Cambridge University Press
Lowe, E. J. (2001). Identity, composition and the self. In Corcoran, K (Ed.), *Soul, body and survival* (pp. 139–58): Ithaca: Cornell University Press.
Lowe, E. J. (2009). Serious endurantism and the strong unity of human persons. In L. Honnefelder, E. Runggaldier, & B. Schlick (Eds.), *Serious endurantism and the strong unity of human persons* (pp. 67–82). Berlin: Walter de Gruyter.
Lupton, D. (2014). *Self-tracking cultures: towards a sociology of personal informatics.* Paper presented at the Proceedings of the 26th Australian Computer-human interaction conference on designing futures: The future of design.
Lupton, D. (2016). *The quantified self*. Wiley.
Luria, A. R., & Vygotsky, L. S. (1992). *Ape, primitive man and child: Essays in the history of behaviour*. New York: Simon and Schuster.
Malafouris, L. (2008a). *At the potter's wheel: An argument for material agency Material agency* (pp. 19–36). New York: Springer.
Malafouris, L. (2008b). Between brains, bodies and things: Tectonoetic awareness and the extended self. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1499), 1993–2002.
Malafouris, L. (2013). *How things shape the mind: A theory of material engagement*. Cambridge: MIT Press.
Marcus, G. (2008, December 18). What if HM had a blackberry? Coping with amnesia, using modern technology. *Psychology Today*.
McCullough, M. (1996). *Abstracting craft: The practiced digital hand*. Cambridge, MA: The MIT Press.
Menary, R. (2007). *Cognitive integration: Mind and cognition unbounded*.

Metzinger, T. (2004). *Being no one: The self-model theory of subjectivity*. Cambridge, MA: Bradford Book.
Metzinger, T. (2009). *The Ego tunnel: The science of the mind and the myth of the self*. New York: Basic Books.
Olson, E. T. (1999). *The human animal: Personal identity without psychology*. Oxford University Press.
Parfit, D. (1984). *Reasons and persons*. Oxford: OUP.
Schechtman, M. (1990). Personhood and personal identity. *Journal of Philosophy, 87*, 71–92.
Schechtman, M. (2005). Personal identity and the past. *Philosophy, Psychiatry, & Psychology, 12*(1), 9–22.
Schechtman, M. (2011). The narrative self. In Gallagher, S (Ed.), *The Oxford Handbook of the Self* (pp. 394–416): Oxford: Oxford University Press.
Schneider, S. & Corabi, J. (this volume). Cyborg Divas and Hybrid Minds.
Schneider, S. (2013). Non-reductive physicalism and the mind problem. *Noûs, 47*(1), 135–153.
Schneider, S. (2019). *Artificial you: AI and the future of your mind*. Princeton: Princeton University Press.
Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences, 17*(11), 565–573.
Smart, P. R. (this volume). *Predicting me: The route to digital immortality*.
Smart, P. R., Heersmink, R., & Clowes, R. W. (2017). The cognitive ecology of the the internet. In S. J. Cowley & F. Vallée-Tourangeau (Eds.), *Cognition beyond the brain* (2nd ed., pp. 251–282). Cham: Springer.
Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data, 1*(2), 85–99.
Turkle, S. (2007). *Evocative objects: Things we think with*. Cambridge: The MIT Press.
Zahavi, D. (2005). *Subjectivity and selfhood: investigating the first-person perspective*. Cambridge: The MIT Press.

Robert W. Clowes is senior researcher and coordinator of the Lisbon Mind and Reasoning Group at the Universidade Nova de Lisboa, Portugal. His research interests span a range of topics in philosophy and cognitive science, including the philosophy of technology, memory, agency, skills, and the implications of embodiment and cognitive extension for our understanding of the mind and conscious experience. He is particularly interested in the philosophical and cognitive scientific significance of new technologies, especially those involving the Internet and artificial intelligence and how these interact with human agency. His work has appeared in a variety of journals, including *TOPOI*, *Review of Philosophy and Psychology*, *AI & Society*, *Phenomenology and the Cognitive Sciences*, *Philosophy and Technology*, and the *Journal of Consciousness Studies*. He received his PhD from the University of Sussex.

Klaus Gärtner studied philosophy at the University of Regensburg. He obtained his PhD at the Instituto da Filosofia da NOVA (Universidade Nova de Lisboa). Currently, he is a researcher at the Departamento de História e Filosofia das Ciências and member of the Centro de Filosofia das Ciências da Universidade de Lisboa in the Faculdade de Ciências da Universidade de Lisboa. He is also a founding member of the Lisbon Mind and Reasoning Group. His research interests include philosophy of mind and cognitive science, philosophy of science, epistemology, and metaphysics.

Chapter 9
Predicting Me: The Route to Digital Immortality?

Paul Smart

To die,—to sleep,
To sleep! perchance to dream…
For in that sleep of death what dreams may come…

—Hamlet, William Shakespeare

9.1 Introduction

Do you want to live forever? If so, then the twenty-first century may be the perfect time to die. Digital immortality has long been a source of fascination for the transhumanist movement, yielding many proposals as to how a given individual might be 'serialized' to a digital medium and then 'resurrected' as part of some digital afterlife. Such accounts have often been the target of philosophical criticism, inspiring more in the way of philosophical invective than they have technological innovation. But is all this about to change? Recently, there has been a renewed interest in the notion of digital immortality, with an increasing number of companies now offering some form of digital afterlife. "Become virtually immortal," reads the slogan of one such company, Eternime.[1] It is, of course, unclear whether this particular commercial offering relates to a virtual form of immortality or a form of immortality that is virtually possible. But notwithstanding these ambiguities, the long-term aims of the company are relatively clear:

> Eternime collects your thoughts, stories and memories, curates them and creates an intelligent avatar that looks like you. This avatar will live forever and allow other people in the future to access your memories.

[1] See http://eterni.me/ (accessed: 7th March 2018).

P. Smart (✉)
Electronics and Computer Science, University of Southampton, Southampton, UK
e-mail: ps02v@ecs.soton.ac.uk

R. W. Clowes et al. (eds.), *The Mind-Technology Problem*, Studies in Brain and Mind 18, https://doi.org/10.1007/978-3-030-72644-7_9

Discussions of digital immortality typically go hand-in-hand with an appeal to some form of *mind uploading*. According to Goertzel and Ikle' (2012), mind uploading is:

> …an informal term referring…to the (as yet hypothetical) process of transferring the totality or considerable majority of the mental contents from a particular human brain into a different substrate, most commonly an engineered substrate such as a digital, analogue or quantum computer. (Goertzel and Ikle' 2012, p. 1)

Perhaps unsurprisingly, there are many different proposals as to how this rather vague objective might be realized. Most proposals advocate the use of advanced technology to record information about the structure of an individual's biological brain. Hayworth (2012), for example, discusses how an advanced imaging technique (Focused Ion Beam Scanning Electron Microscopy) might be used to map the structure of whole brain neural circuits, yielding a more-or-less complete model of the human connectome—i.e., the connection matrix of the human brain (Sporns et al. 2005). Inasmuch as it is this structural description of the brain that defines who and what we are (see Seung 2012), then such approaches have an obvious appeal, even if they are still recognized as being impractical or beyond the limits of current technology.

The present chapter describes an approach to digital immortality that is similar, at least in spirit, to many forms of mind uploading. Where it departs from previous accounts is with respect to the approach taken to model (and re-generate) the functional dynamics of the human brain. Instead of trying to directly map the detailed microstructure of the biological brain using imaging or tracing techniques, the present approach is rooted in the use of machine learning techniques, especially those forms of machine learning whose styles of computation, representation, and overall architecture share some similarity with recent models of brain-based (or at any rate, cortical) processing. The inspiration for this approach is based on a recent neurocomputational model of brain function that depicts the biological brain as a hierarchically-organized predictive processing system, constantly engaged in the attempt to predict its own activity[2] at a variety of (increasingly abstract) spatial and temporal scales (Clark 2016, 2013b; Friston 2010). This account of brain function has been the target of considerable scientific and philosophical interest, at least in part because the account is deemed to be relevant to a broad swath of seemingly disparate psychological phenomena, including learning, attention, perception, action, emotion, imagination, memory, and various forms of mental illness (Clark 2016; Friston et al. 2014). Beyond this, however, the vision of the brain as a hierarchically-organized predictive processing system is one that is reflected in recent approaches to machine learning, especially those associated with deep learning systems (Bengio 2009; LeCun et al. 2015). Such forms of convergence are interesting given the recent successes of deep learning on a number of Artificial Intelligence (AI) problems, and they may even mark the beginnings of a still

[2] In essence, the brain is viewed as a multi-layered prediction machine, with 'higher' layers attempting to predict the activity of 'lower' layers. It is in this sense that the brain can be seen to predict its own activity, i.e., to predict the activity of its constituent neural elements.

somewhat ill-defined path towards experientially-potent forms of machine cognition. The aim of the present chapter is to describe these two areas of research (i.e., predictive processing models of brain function and deep machine learning) with a view to outlining an approach to digital immortality that highlights the relevance of research into deep learning, virtual reality, and big data processing. The technologies associated with these research areas are likely to have a substantial impact on various spheres of human activity throughout the twenty-first century.

The chapter is structured as follows: Section 9.2 outlines the broad shape of the predictive processing (PP) framework, as discussed by Clark (2016), Friston (2010), and others. It focuses on a key feature of the PP account, namely, the use of generative models to construct the sensory signal 'from the top down'. Section 9.3 aims to draw attention to some of the similarities between the PP account of brain function and deep learning systems. Section 9.4 then goes on to suggest that the link between deep learning systems and PP serves as the basis for a particular approach to digital immortality: one that is rooted in the idea that deep learning systems might be used to recreate the generative models that are acquired by biological brains as a result of prediction-oriented learning. Section 9.5 reflects a shift in focus, from machine learning to big data. It discusses some of the problems confronting the approach to digital immortality presented in Sect. 9.4. In particular, Sect. 9.5 raises questions about the sort of data that ought to be collected, as well as some of the technical, social, and ethical challenges that are likely to confront the data collection effort. Section 9.6 explores the role of virtual reality technologies in establishing some sort of digital afterlife, with a particular emphasis on the notion of embodiment. Finally, Sect. 9.7 concludes the chapter.

9.2 Predictive Processing

Over the course of the past decade, a particular view of the brain has become increasingly popular, both in cognitive neuroscience and the philosophy of mind. This is a view that sees the biological brain as a hierarchically-organized system that is constantly striving to predict its own internal activity, relative to the play of energy across the organism's sensory surfaces (Clark 2016). The most popular version of this account is known as the predictive processing account of cognition (PP for short).

Some insight into the general flavor of PP can be gleaned by comparing the PP approach to perception with its more traditional counterpart (see Fig. 9.1). On the traditional view, perception occurs via the stepwise analysis of incoming sensory information, with more abstract features being detected at progressively higher levels of the cortical hierarchy (see Fig. 9.1a). The aim, in this case, is to analyze the upward/forward-flowing stream of information to the point where action can be coordinated with respect to abstract properties of the perceptual scene.

The PP view of perceptual processing is somewhat different (see Fig. 9.1b). Here, the stream of incoming sensory information is met with a downward/

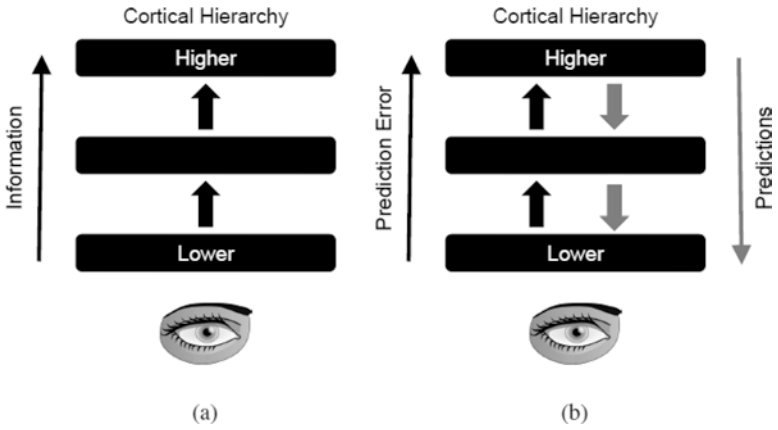


Fig. 9.1 Two approaches to perceptual processing. (**a**) The traditional approach to perceptual processing is characterized by an upward/forward-flow of information through a succession of cortical regions. (**b**) This contrasts with the PP approach, which emphasizes the role of backward/downward-flowing predictions in suppressing the upward/forward flow of information

backward cascade of predictions that emanate from progressively higher layers of the processing hierarchy. The purpose of this downward-flowing stream is to predict the activity of neural circuits at each layer in the hierarchy, with the forward/upward-flowing stream of information being used to communicate the mismatch between actual and predicted activity. The overall aim of the system, in this case, is to minimize prediction error and thus suppress the forward flow of information. From an information-theoretic standpoint, prediction error is seen to provide a measure of "free energy," which is defined as the "difference between an organism's predictions about its sensory inputs (embodied in its models of the world) and the sensations it actually encounters" (Friston et al. 2012, p. 1). Reductions in prediction error therefore correspond to reductions in free energy, which, over the longer term, equates to a form of entropy minimization (see Friston 2010). As noted by Clark (2016):

> …good [predictive] models…are those that help us successfully engage the world and hence help us to maintain our structure and organization so that we appear—over extended but finite timescales—to resist increases in entropy and (hence) the second law of thermodynamics. (Clark 2016, p. 306)

The means by which predictive capabilities are acquired by the brain is often depicted as a form of perceptual learning. In essence, prediction error is seen to promote changes in synaptic strength that reconfigure the structure of neural circuits, enabling higher-level neural regions to better predict the activity of lower-level regions. Crucially, one of the upshots of this particular form of prediction-oriented learning is the installation of a hierarchically-organized model that (by virtue of the organism's sensory contact with the world) tracks the hidden causes (or latent variables) that govern the statistical structure of incoming sensory information. An important feature of these models is that they are *generative* in nature. That is to say, the models encoded by the neural circuits at each layer in the

hierarchy must be such as to allow each layer to predict activity in the layer below. This means, in effect, that each layer is able to generate the information encoded by the lower layer, and this extends all the way out to the lowest levels of the hierarchy, i.e., to the point where the torrent of downward-flowing predictions meets the incoming tide of sensory information.

It is widely assumed that the overall result of this hierarchical organization is a multi-layer generative model that embodies the causal structure of the environment. In particular, it is assumed that by virtue of the attempt to recapitulate the activity of lower levels, and thus accommodate the incoming sensory signal, the brain is attempting to model the interacting set of worldly causes that give rise to particular kinds of sensory stimulation. In a sense, successful prediction is like a form of 'understanding', where the understanding in question concerns the causal forces and factors that shape whatever bodies of sensory information exist within the organism's local environment (see Clark 2013a). It is at this point that the commitment to multi-layer, hierarchically-organized neural architectures takes on a special significance, for it seems that such an organization is ideally suited to the kind of world in which we humans live—a world built around a structured nexus of interacting, and often deeply nested, causal forces. The goal of perception, according to PP, is to invert this casual structure by using a generative model to infer the causes of sensory input (see Fig. 9.2). "The hierarchical structure of the real world," Friston (2002) suggests, "literally comes to be reflected by the hierarchical architectures trying to minimize prediction error, not just at the level of sensory input but at all levels of the hierarchy" (Friston 2002, pp. 237–238).

As noted by Clark (2016), generative models may lie at the heart of a number of cognitive phenomena, including our capacity for hallucination, dreaming, fantasy, and the potential for self-generated forms of mental imagery. In acquiring a

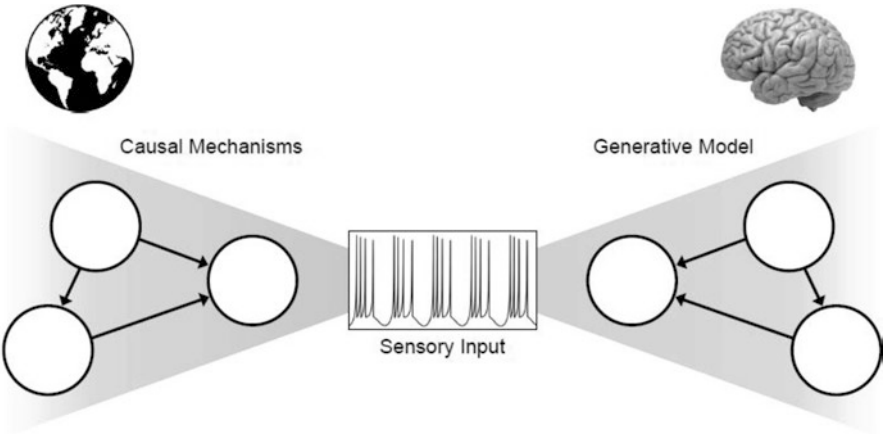


Fig. 9.2 A generative model describes how variables or causes in the environment conspire to produce sensory input. Perception then corresponds to the inverse mapping from sensations to their causes

generative model, the brain develops a capacity to drive patterns of neural activity from the top-down, and in doing so it is able to recreate the patterns of activity that would be instantiated if the organism were to be confronted with a particular pattern of sensory input. "Probabilistic generative model based systems that can learn to visually perceive a cat," suggests Clark (2013b), "…are, ipso facto, systems that can deploy a top-down cascade to bring about many of the activity patterns that would ensue in the visual presence of an actual cat" (Clark 2013b, p. 198). This seems to be of profound importance relative to our understanding of many aspects of our mental lives. Generative capacities thus seem relevant to our ability to dream the non-existent, imagine the counterfactual, anticipate the future, and (via memory) reconstruct the past.

In addition to causing us to rethink the relationship between seemingly distinct cognitive phenomena, the PP account also informs our view of what is perhaps the most elusive part of our cognitive economy: conscious experience. In particular, the appeal to generative models is sometimes seen to reinforce an approach to consciousness that portrays it as a form of virtual reality (see Revonsuo 1995) or controlled hallucination. Metzinger (2003), for example, suggests that a fruitful way of looking at the brain is to view it:

> …as a system which, even in ordinary waking states, constantly hallucinates at the world, as a system that constantly lets its internal autonomous simulational dynamics collide with the ongoing flow of sensory input, vigorously dreaming at the world and thereby generating the content of phenomenal experience. (Metzinger 2003, p. 52)

As a means of establishing a better grip on this idea of phenomenal experience as a form of controlled hallucination or actively constructed virtual reality, it may help to consider a hypothetical situation in which the predictability of the environment (or the predictive power of the brain's generative model) is such that the activity at each layer in the neural processing hierarchy is perfectly predicted by the layer above. In such a situation, the incoming flow of sensory information is met by a cascade of downward-flowing predictions, which perfectly captures the activity at each and every layer of the processing hierarchy. Such a state-of-affairs is interesting, for the predictive successes of each layer eliminate any forward (or upward) flow of information (i.e., from lower to higher layers in the hierarchy). According to the PP account, recall, the forward flow of information corresponds to prediction error—the mismatch between actual and predicted activity at each level in the hierarchy. But in this particular case—let us call it the *no-prediction-error-case*—there is no prediction error, and thus there is no forward/upward flow of information. What we are left with, therefore, is a purely backward or downward flow of information, from higher cortical regions all the way out to the point of sensory input.

Of course, actual instances of the no-prediction-error-case are unlikely to be found in the real world. Real-world environments are seldom perfectly predictable, and the activity of neural circuits is often characterized by the presence of neuronal noise (e.g., Faisal et al. 2008). Thus even if organisms were to seek out a dark, unchanging chamber as per the worries raised by the "Dark-Room Problem" (Friston et al. 2012), it is unlikely that such organisms would be completely free of

prediction error. As a simple thought experiment, however, the no-prediction-error-case is useful in helping us get to grips with the idea of perceptual experience (and perhaps phenomenal experience, more generally) as something that is generated from the 'inside out'. Assuming that perceptual experience occurs as a consequence of the formation of stable neural states (i.e., those that successfully predict the activity of other neural regions and are thus unperturbed by any form of prediction error), then the no-prediction-error-case is one in which the experience of (for example) seeing a scene occurs in the absence of any forward flow of information through the brain. The result is a view of conscious experience as something that is actively generated by the brain as part of its attempt to model the causal structure of the sensorium and thus predict its own (sensorially-shaped) neural activity.

The upshot of all this is a vision of the brain as a form of virtual reality generator—the biological equivalent of technologies that render virtual objects and virtual worlds. According to this vision, aspects of our daily conscious experience are tied to the brain's attempt to acquire and deploy generative models that track the causal structure of the external environment. In particular, it seems that conscious experience might be linked to the activation of representations corresponding to an interacting set of external causes that are acquired as the result of an attempt to predict the structure of incoming sensory information.

This is a compelling, although still somewhat puzzling, vision. It is a vision that depicts our phenomenal experience as something akin to a simulation of reality, and it is a vision that blurs the distinction between ostensibly distinct cognitive phenomena, such as perception, imagination, dreaming, and fantasy. Relative to such a vision, it is perhaps easy to think of life as nothing more than a dream. To echo the views of Metzinger, what we call waking life may be nothing more than a form of "online dreaming" (Metzinger 2003, p. 140).

9.3 Dream Machines

The PP account bears some interesting similarities to recent work in machine learning, particularly that which focuses on so-called deep learning systems (Bengio 2009; LeCun et al. 2015). As with the PP model of brain function, deep learning emphasizes the importance of multi-layer architectures, with 'higher' layers yielding more abstract representations of the response patterns exhibited by 'lower' layers (at least in some systems). The notion of generative models also marks a point of commonality between PP and at least some strands of deep learning research. Here, the attention of the machine learning community has shifted away from a traditional focus on the discriminative capacities of neural networks (e.g., their ability to discriminate between objects of different types) towards a better understanding of their generative capabilities (i.e., their ability to re-create bodies of training data).

This shift in focus—from discriminative to generative capacities—is important, for it highlights the potential relevance of deep learning systems to debates about

digital immortality. Inasmuch as deep learning systems are able to emulate the functionality of the brain with respect to the acquisition and deployment of generative models (and inasmuch as generative models are revealed to be a cornerstone of the human cognitive economy), then we might wonder whether a deep learning system could be used to re-create the generative models embodied by a biological brain. Perhaps if we could capture the streams of sensory data against which brain-based generative models take shape, we could then use this data to train a deep learning system and thereby reinstate (some of) the cognitive properties of a given human individual. This is the essence of an approach to digital immortality that highlights the potential relevance of two key twenty-first century technologies—deep learning systems and big data technologies—to issues of digital immortality.[3] We will explore some of the implications (and problems) associated with this approach in subsequent sections. For now, however, let us direct our attention to the nature of the (putative) link between deep learning systems and the PP account of brain function.

When it comes to deep learning systems with generative capacities, a number of systems have been the focus of recent research attention. These include Deep Belief Networks (DBNs) (Hinton 2007a, b), variational autoencoders, and Generative Adversarial Networks (GANs) (Goodfellow et al. 2014). DBNs are a particular kind of deep learning system. They are composed of multiple layers of what are called Restricted Boltzmann Machines (RBMs). These RBMs are a type of neural network consisting of two layers: a visible layer and a hidden layer. The nodes within each layer are connected to nodes in adjacent layers; however, there are no intra-layer connections (i.e., nodes within a particular layer are not connected to nodes in the same layer).[4] The nodes in the visible layer represent the data that is presented to the RBM, and the goal of the hidden layer is to capture higher-order correlations between the data that is represented at the visible layer. Typically, all the nodes in the RBM are binary, with two states represented by the digits '0' and '1'. This means that in cases where the RBM is presented with a black and white image, the nodes at the visible layer will represent the image using a binary data vector whose elements represent the individual pixel intensities of the image (e.g., '1' for white and '0' for black). Relative to this case, the nodes in the hidden layer now function as binary feature detectors that seek to model the higher-order correlations between pixel values at the visible layer. In particular, the aim during learning is to configure the weights associated with the top-down connections (from hidden layer to visible layer) such that the hidden layer is able to recreate the training data represented by

[3] It is doubtful whether the current state-of-the-art in deep learning is sufficient to achieve the sort of digital immortality vision being proposed here. Nevertheless, it is worth bearing in mind that deep learning research is likely to be a prominent focus of global research attention over the next 10–20 years. Given the amount of time, effort, and money that is likely to be devoted to deep learning systems in the coming years, it is likely that we will see significant changes in their capabilities during the course of the twenty-first century.

[4] This highlights one of the differences between DBNs and PP accounts of cognition. In PP, it is typically assumed that elements within the same layer of the processing hierarchy engage in some form of lateral processing.

the nodes of the visible layer. It is in this sense that the RBM's model of the training data is said to reside in its top-down connections.

Clearly, a single layer of binary feature detectors is unlikely to capture all the latent structure that exists within a complex set of images, especially when we reflect on the complexity of the interacting causal processes that conspire to generate individual pixel intensities (see Horn 1977). For this reason, additional layers are added to a base RBM to expand its representational capabilities. It is at this point that a single, two-layer RBM begins to morph into a multi-layer DBN (see Fig. 9.3a). As each layer is added, the new layer is treated as the hidden layer of a new RBM, while the erstwhile hidden layer of the original RBM now functions as the visible layer. In effect, the representations of the original hidden layer now become the training data (or 'sensory' input) for the new layer, and the goal of the new layer is to learn a suite of more abstract representations that capture the dynamics of the layer below. Perhaps unsurprisingly, the addition of each new layer enhances the system's ability to model abstract structural regularities, thereby improving its capacity to generate the training data at the lowest layer in the hierarchy (i.e., the visible layer of the original RBM).

The upshot of this kind of (incrementally-oriented) learning regime is a multi-layer neural network that is, in effect, a composite of multiple RBMs (see Fig. 9.3b). This system is what is typically dubbed a DBN. At this point, the system possesses a good generative model of the target domain, as represented by the training data,

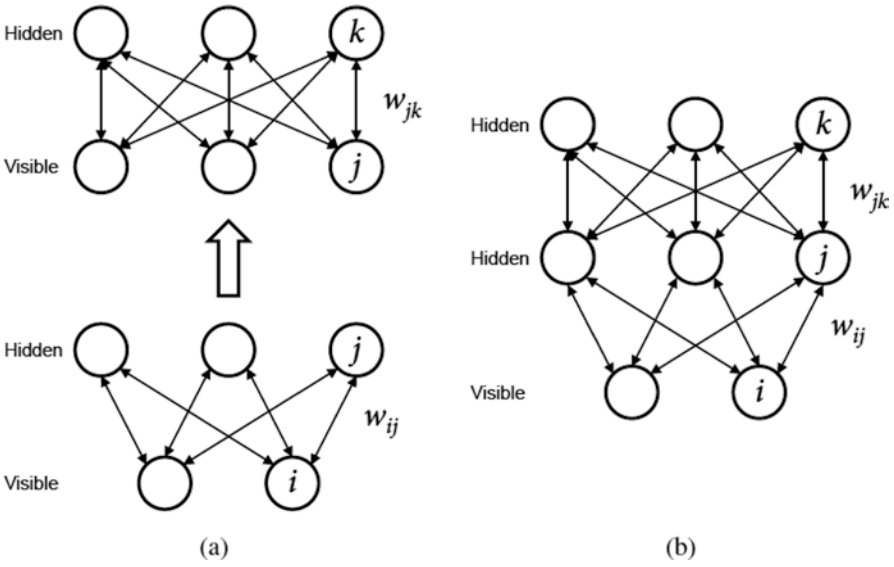


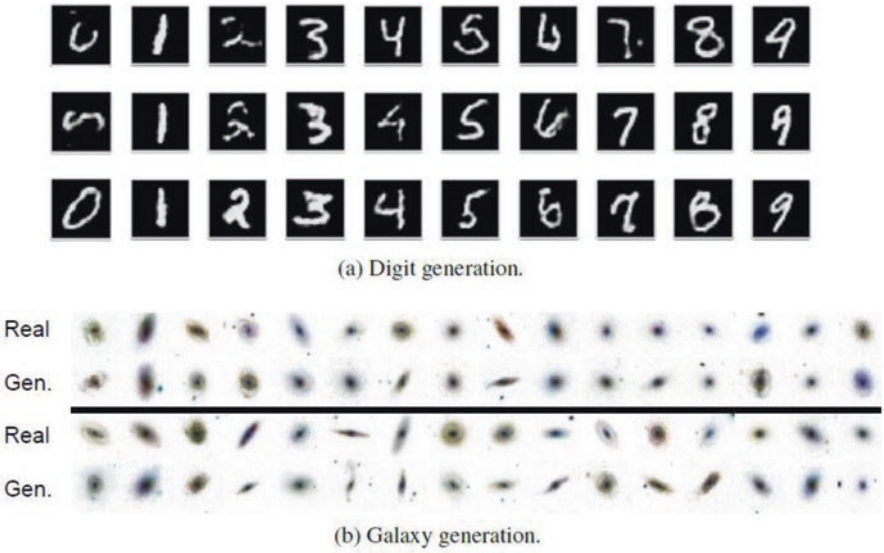
Fig. 9.3 Restricted Boltzmann Machines and Deep Belief Networks. (**a**) Two separate RBMs. The stochastic binary variables in the hidden layer of each RBM are symmetrically connected to the stochastic binary variables in the visible layer. There are no connections within a layer. The higher-level RBM is trained using the hidden activities of the lower RBM as data. (**b**) A DBN formed from the merger of two RBMs

but it isn't necessarily well-suited to other kinds of tasks, such as the classification of images into particular classes. Nevertheless, in being forced to recreate the training data, the network has learned a great deal about the hidden causes or latent variables that structure the training data. What the network has learned, in effect, is a way of 'explaining' each input vector (each sample of sensory data) in terms of a nexus of interacting and deeply nested (hidden) causes, where the notion of a good explanation corresponds to "a binary state vector for each layer that is both likely to cause the binary state vector in the layer below and likely to be caused by the binary state vector in the layer above" (Hinton 2010, p. 179). It turns out that this 'explanatory' capability serves as the basis for enhanced performance in a variety of task contexts. In the case of Hinton's early work with DBNs, for example, a DBN was trained with images of handwritten digits taken from the MNIST data set. The DBN's subsequent discriminative performance was tested by extending the network with a set of 'label' nodes corresponding to the kind of conceptual distinctions that we humans make when dealing with the realm of handwritten digits (i.e., our ability to recognize a particular image as representing a particular number). A variety of studies have shown that this kind of approach—essentially treating digit classification as something of a post-processing step relative to the primary goal of acquiring a generative model—is able to deliver superior performance compared to networks that are trained with conventional back propagation techniques and random initial weights (Hinton and Salakhutdinov 2006).

In addition to highlighting the performance benefits of DBNs when it comes to the recognition of handwritten digits, Hinton's work also provides a compelling demonstration of the generative capacity of such architectures (see Hinton 2007b). With the addition of a set of label nodes representing digit types, Hinton was able to selectively activate particular label nodes and then observe the data vector (the sensory output) produced by the network at the visible layer. Figure 9.4a illustrates some of the images generated using this method.

A further demonstration of the generative capacity of deep learning systems comes from a study by Ravanbakhsh et al. (2017). The purpose of Ravanbakhsh et al.'s (2017) study was to generate images of galaxies for the purpose of calibrating astronomical equipment. As part of the study, Ravanbakhsh et al. exposed two kinds of deep learning system—namely, variational autoencoders and GANs—to images of real galaxies taken from the Galaxy Zoo data set.[5] As a result of training with respect to these images, the deep learning systems acquired a generative model of the target domain, enabling them to produce galaxy images similar to those contained in the real-world data set (see Fig. 9.4b). In a sense, of course, these images are 'fakes', since the galaxies they represent do not exist. At the same time, however, there is surely something compelling about the generative abilities exhibited

[5] The Galaxy Zoo data set consists of 900,000 galaxy images, which were collected as part of the Sloan Digital Sky Survey. The data set was originally used as part of a citizen science project investigating the distribution of galaxies with particular morphologies (see Lintott et al. 2008). It has since been used as the basis for a number of studies exploring the capacities of both conventional and deep neural networks (Dieleman et al. 2015; Banerji et al. 2014).



(a) Digit generation.

(b) Galaxy generation.

Fig. 9.4 Examples of sensory data generated by deep learning systems. (**a**) The output from a generative model trained with handwritten digits. Each row shows 10 samples from the generative model with a particular digit label clamped on. The top-level associative memory is run for 1000 iterations of alternating Gibbs sampling between samples (source: http://www.cs.toronto.edu/~hinton/digits.html). (**b**) Actual ('Real') versus generated ('Gen.') galaxy images. Actual images are taken from the Galaxy Zoo data set; generated images were created using a conditional generative adversarial network. (Source: Ravanbakhsh et al. 2017)

by these systems. Their abilities remind us of our own human ability for creativity—our ability to use our daily waking experience as the point of departure for excursions into a realm of purely imagined objects, ideas, and other fantastical constructions. As was mentioned above, it is here that recent work into generative models provides us with a potential insight into what is perhaps the most elusive aspect of the human cognitive economy—namely, our capacity to imagine, to fantasize, and to dream. We may, of course, never know whether deep neural networks should count as *bona fide* 'dream factories'—systems that are genuine dreamers and imagineers—but we do know that they (like us) are creative engines—systems capable of using high-level abstract representations for the purpose of creating digital artefacts (e.g., images). This should perhaps give us pause when we wonder (and sometimes worry) about the capacity of AI to encroach on the oft cherished realm of human creativity. For perhaps the early successes of generative deep learning systems intimate at something more profound: an ability to emulate (and perhaps even surpass) our best intellectual triumphs and artistic accomplishments. If a deep neural network can generate an image of a galaxy (which is, after all, just a sequence of binary digits), then why should it not be able to create other kinds of digital artefact, such as a 3D object, a movie, or the design for a deep neural network capable of delivering state-of-the-art advances in machine-based generative capabilities?

9.4 Generating Me

It should by now be clear that the representational and computational profile of at least some forms of deep learning system echo those that lie at the heart of the PP approach to cognition. There are, of course, important differences. Deep learning systems seldom feature any form of intra-layer processing and important elements of the PP model (e.g., real-time prediction error estimation and the precision-weighted modulation of information flows) are absent (or at least under-represented) in current deep learning systems. Nevertheless, the basic commitment to hierarchically-organized processing schemes and the emphasis on generative capabilities has paid substantial dividends, yielding important advances in problem areas that have long stymied the efforts of the AI community (LeCun et al. 2015; Najafabadi et al. 2015). In addition to notable successes in the areas of language processing and computer vision, deep learning research has yielded interesting results in a number of more 'niche' areas, helping us understand the generative mechanisms responsible for the structure of neural circuits (Betzel et al. 2016) and providing us with predictive models of drug molecule activity (Ma et al. 2015). The question, of course, is whether these early successes help to reveal the beginnings of a path that makes inroads into the problem of digital immortality.

It is at this point that I will opt to bite the bullet, so to speak, and suggest that some form of deep learning system—specifically, a system whose computational profile is closely aligned with the PP model—is, indeed, a technology that ought to be considered by the proponents of digital immortality. The basic idea is that a particular form of deep learning system—let us call it a Synthetic Predictive Processing (SPP) system to emphasize the overlap with the PP model—is tasked with the objective of acquiring generative models that resemble those acquired by an individual's biological brain. The idea, in essence, is for a SPP system to exist as a form of constant digital companion that accompanies the individual across the course of their life and participates in a form of lifelong learning. The ultimate goal of such a system is to deal with the same sort of predictive challenges that confront the biological brain. The hope is that in dealing with these challenges, the SPP system will come to acquire a set of probabilistic generative models whose generative capacities are functionally similar to those acquired by its biological counterpart. Inasmuch as our 'online dreams' are produced by generative models in the biological case, is there any reason to think that the general shape of those dreams could not be manufactured by a synthetic predictive processing machine with more-or-less the same generative capabilities?

There are multiple possible variations on this sort of idea. The proposal just outlined is what I will dub the *first-person proposal*. It seeks to situate a SPP system within the same sensory environment as that in which the biological brain is itself embedded. An alternative proposal comes in the form of what I will call the *third-person proposal*. In this case, the human agent is the primary target of prediction-oriented learning. The goal is thus to monitor the responses of the human individual and predict those responses relative to features of the individual's local

environment. The upshot, perhaps, is a generative model that embodies something about the character of an individual—their propensity to act in particular ways in certain situations, their tendency to spend their hard-earned money on certain commodities, and their likely linguistic responses to certain kinds of conversational context.

These two proposals differ with respect to the kind of 'sensory' data over which generative models are formed. In both cases, however, we see a commitment to uncovering the deep structuring causes or latent variables that best explain the shape of the evolving sensory signal. Of the two proposals, the first-person proposal is likely to be the one that best serves the interests of the digital immortality agenda. But this does not mean that the third-person proposal is entirely without merit. It is, indeed, the third-person proposal that probably best reflects the current interests of the technological community in preserving some trace of an individual (e.g., by maintaining repositories of an individual's social media posts). Crucially, in thinking about the *kind* of data that can be used to shape the dynamical profile of PP systems, we are provided with some insight into the shape of what might be called the 'morphospace of the mind' (i.e., the universe of all possible minds) (see Mitteroecker and Huttegger 2009),[6] as well as (perhaps) additional means of achieving digital immortality. Consider, for example, the way in which current deep learning systems are being used to create generative models of the human connectome (Betzel et al. 2016), or the way in which so-called 'brain reading' devices are yielding new opportunities to model the real-time response profile of the biological brain using prediction-oriented learning and generative modelling techniques (see van Gerven et al. 2010).

9.5 Memories for Life (and Beyond)

The form of digital immortality envisioned in the previous section relies on the availability of substantial bodies of data about the individual. From this perspective, the process of 'generating me' emerges as a particular form of big data challenge—one that dovetails with the interests and concerns of a number of research communities. In yielding the data for prediction-oriented learning processes, it should thus be clear that the vision of digital immortality scouted above establishes a natural point of contact with the emerging discipline of data science (Committee on the Analysis of Massive Data, 2013), especially when it comes to the acquisition, analysis, and

[6] The general idea, here, is that different kinds of data environment provide the basis for different kinds of minds, with phenomenological differences linked to the causal mechanisms that operate in each environment. A generative model of the human social environment, for example, might yield a 'mind of society' that tracks the hidden causal structure of social mechanisms. Inasmuch as subjective experiences are tied to the properties of generative models, then such a mind may yield a subjective reality that is profoundly different from the sort of 'reality' we know (or could, perhaps, even imagine).

storage of big data. Of particular interest is work that seeks to use big data for the purpose of enhanced prediction, typically by drawing on a capacity for predictive modelling (Dhar 2013). Another prominent point of interest concerns the application of deep learning methods to big data assets. As noted by Najafabadi et al. (2015), the ability to detect statistical regularities in data using unsupervised learning techniques makes deep learning particularly well-suited to dealing with some of the analytic challenges associated with big data science.

It has to be said, of course, that the data sets targeted by contemporary data science are unlike those that form the basis of the aforementioned first-person proposal. For the most part, the term "big data" is often applied to bodies of data pertaining to scientific observations (e.g., astronomical data) or data that tracks the properties of *multiple* human individuals (e.g., epidemiological data). There is, however, a growing interest in the analysis of data that is gleaned from individual human agents. Indeed, the analysis of individual (or personal) data forms the basis of research into so-called *personal informatics*, which is being driven, at least in part, by the availability of new digital recording devices, such as wearable cameras, smartphones, and activity monitors. The use of technology to record or track information about individual human subjects is also a central element of work that goes under the heading of the *quantified-self* (Lupton 2013; Swan 2013),[7] which has highlighted the way in which self-tracking technologies can be used to record data relating to (for example) body weight, energy levels, time usage, heart rate, body temperature, exercise patterns, sleep quality, sexual activity, diet, dreams, and blood chemistry. It is at this point, perhaps, that some of the data-oriented challenges associated with the aforementioned vision of digital immortality—specifically those pertaining to the acquisition of training data—look to be a little less formidable. There is no doubt a sense in which our current modes of self-related data acquisition remain deficient relative to the requirements of the digital immortality vision. But there is, I suggest, no reason to doubt the overall feasibility of the data tracking effort: our current technologies are already yielding ample data about a range of physiological and behavioral variables, and such data is already recognized as a valuable source of information, with the use of machine learning and other forms of big data analysis poised to reveal hidden structure in our self-generated digital trails (Fawcett 2015; Phan et al. 2017; Hoogendoorn and Funk 2018).

Future tracking technologies are likely to expand both the volume and variety of data that can be acquired from individuals during the course of their lives, altering the opportunities for prediction-oriented learning and the acquisition of generative models. When it comes to our ability to monitor individual movements, for example, research into so-called artificial skin (or e-skin) devices (Yokota et al. 2016; Someya et al. 2004) and smart fabrics (Foroughi et al. 2016; Wang et al. 2014) is likely to be particularly significant. In supporting the acquisition of data about physical movement, such technologies may provide a means to track information of a

[7] A variety of other terms are sometimes used to refer to the same phenomenon. These include "lifelogging," "measured me," "self-tracking," and "self-surveillance."

proprioceptive nature. This looks to be important inasmuch as we see the route to digital immortality as predicated on the attempt to recapitulate the sorts of predictive processing undertaken by the biological brain. From this perspective, digital recording technologies function as a substitute for biological sensors, recreating the kind of information streams that characterize the sensorium of the biological individual. In this respect, there seems to be ample cause for optimism. In addition to the monitoring of information of a (broadly) proprioceptive nature, technological advances are likely to improve our ability to record information from both the corporeal and extra-corporeal environment (i.e., information of a broadly interoceptive and exteroceptive nature). The advantage of this particular approach to data acquisition is that it depicts a SPP system as in more or less the same position as the biological brain. Both the brain and its synthetic counterpart are thus attempting to establish a predictive grip on common bodies of sensory information, and they are thus under more-or-less the same pressure to acquire and deploy similar generative models. It is this particular approach—the approach mandated by the first-person proposal in Sect. 9.4—that is perhaps best placed to deliver the most potent forms of digital immortality. For the goal of digital immortality is not merely to learn about a specific individual as an object of study (as per the third-person proposal); it is rather to duplicate the generative structures that make a particular individual the person they are. Inasmuch as we see the elements of the self—our memories, our personalities, our hopes, our fears, and our dreams—as inhering in the structure of a complex multi-layer network that is progressively shaped by our contact with the sensorium, then the quest for digital immortality is perhaps best served by presenting SPP systems with the same bodies of sensory data that confront their neurobiological counterparts.

Is any of this remotely feasible? There are, to be sure, plenty of challenges that confront the attempt to record personal data and use it for the purpose of recreating the functional profile of a brain-based generative model. To my mind, the challenge of emulating the biological brain's predictive and generative capabilities is one of the more daunting challenges, and certainly one that is more daunting than the challenge of acquiring large-scale bodies of personal data. Although research into predictive processing and deep learning exhibits a degree of convergence on hierarchical organizations, generative models, and (to a lesser extent) prediction-oriented learning, there remains a somewhat worrying gap in our understanding of how to emulate the representational and computational wherewithal of the biological brain. Inasmuch as progress in this area lags behind our capacity to capture and store the data that will ultimately be used to train future forms of deep learning system, perhaps there is room within the emerging discipline of data science for a program of research devoted to *data cryogenics*—a field of scientific and engineering research that seeks to preserve bodies of digital data until such times as deep learning systems are deemed able to raise the dead. To me at least, this idea seems no less plausible, and no more outlandish, than the forms of resurrection that are envisioned by those advocating conventional forms of cryogenic preservation.

There is no doubt a further worry raised by all this talk of digital tracking and data monitoring—one that is already felt by those who are ever-more intimately

connected to a surrounding penumbra of digital devices. The concern is that the PP route to digital immortality is one that feeds directly into existing fears about digital surveillance and privacy violation. There is clearly a sense in which such fears, at least as they relate to the present analysis, are justified. Personally, I do not doubt that the sort of vision outlined here will require some degree of privacy violation, and I doubt whether technological advances will do much to assuage such fears, especially if such data is to be 'handed' over to third parties for safekeeping. Perhaps this is something that individuals will need to decide for themselves. Ultimately, it may be the case that privacy is just the price we pay for the possibility of everlasting life.

9.6 The Afterlife

Your biological life has come to an end. You have done your best to weave a digital fabric that tracks the defining moments of your existence. You have, you hope, created some nice memories for your SPP system to capture and model, and you hope that in the process of recreating those moments, some part of you will be preserved. There have, of course, been ups and downs, disasters as well as triumphs, brief moments of happiness punctuated with perhaps longer periods of despair. It does not matter now. Your life is over. Time to die.

But is that necessarily the end of the story? Does the departure of your biological body mean that you yourself are gone, irrevocably lost to those you loved and to those who loved you in return? If anything I have said thus far is anywhere near the mark, then it should be clear that the tale is not quite over. There is, it seems, space for a few pages more.

What, then, is the final part of this vision of digital immortality? What happens once the biological body has fulfilled its purpose and been laid to rest? Arguably, no form of digital immortality is complete without a corresponding form of digital resurrection. But what is the nature of this resurrection relative to the present vision of a SPP system that seeks to emulate the generative capabilities of the biological brain? There are, I suspect, many options available here, but I will choose to limit my attention to the role of virtual reality technologies in supporting a digital afterlife.

Consider first the idea that advances in holographic computing could be used to render an individual in holographic form. This idea is perhaps best exemplified by the character, Joi, in the movie *Blade Runner 2049*. Joi is a virtual companion for the movie's main protagonist, K, who is a replicant blade runner. Unlike K, Joi has no substantive physical presence in the world. She is instead a being made of light; a cinematic entity projected *into* an onscreen (physical, albeit fictional) world. There are, of course, no real-world counterparts to Joi at the time of writing—for the time being at least, she exists solely in the realms of fiction and fantasy. There are, however, reasons to think that something akin to a Joi-like entity might be possible once we reach the midpoint of the twenty-first century—once our own timeline coincides with the timeline of the *Blade Runner 2049* universe. In this respect,

it is worth noting the recent progress that has been made in the development of mixed reality devices. One example is the Microsoft HoloLens, which renders virtual objects (called holograms) within the local physical environment of a human user. Other research establishes an even closer alignment with the *Blade Runner 2049* vision. Consider, for example, research into so-called volumetric displays, which render virtual objects as three-dimensional light displays that can be viewed by multiple users (from multiple angles) without the use of headsets or other user-worn technology (e.g., Smalley et al. 2018). Indeed, in some respects, the capabilities of today's holographic technologies have already surpassed that depicted in *Blade Runner 2049*. In the movie, Joi is a character who can be seen but not touched; her status as a hologrammatic entity precludes the possibility of physical contact and this complicates the nature of her relationship with her physical companion, K. Relative to the thematic structure of the movie, of course, Joi's ethereality is important, for it encourages us to reflect (*inter alia*) on the 'reality' of relationships that transcend the physical/virtual divide. In the real world, however, our interactions with holograms may be far less intangible affairs. Recent research has thus already demonstrated the possibility of so-called touchable or haptic holograms—holograms that can not just be seen, but touched, felt, and even moved (see Kugler 2015).

Here, then, is one of the possibilities for a digital afterlife: individuals will be resurrected as hologrammatic entities—entities that 'live' among us as virtual 'ghosts'. These will be beings whose perceptuo-motor exchanges with the real world are driven by whatever generative models were acquired as part of a biological life within that very same world. Such beings will sense the world (via technological sensors) and implement actions within that world (via changes in photonic rendering technology). Whether they will be able to interact with us, in the sense of being able to touch us, remains to be seen We may, however, still be 'moved' by the presence of these virtual souls, even if the more tactile elements of a human relationship should fail to survive the transition to holographic 'heaven'.

Of course, one way of dealing with the problems thrown up physical reality is to retreat from it altogether. Perhaps, then, a second possibility for the digital afterlife is to situate an individual's SPP system within a purely virtual environment, similar to those built around the use of contemporary game engines. This scenario will probably require little in the way of an introduction, for the idea of a life within a virtual (sometimes self-created) world is one that has been explored by a number of cultural products. Movies such as *The Matrix* and *The Thirteenth Floor* deal with the more general notion of life within a purely virtual environment, while the post-mortem possibilities of virtual environments are explored by the movie *Vanilla Sky* and (my personal favorite) the *San Junipero* episode of the *Black Mirror* sci-fi series.

Both these scenarios rely on the use of virtual reality technologies to address the challenges posed by the demise of an individual's biological body.[8] The technological

[8] These scenarios do not, of course, exhaust the possibilities for digital resurrection. In addition to virtual reality technologies, the twenty-first century is likely to see significant advances in the development of biomimetic materials, 3D printing technology, and robotic systems. These may

challenges associated with these scenarios are no doubt immense, but the appeal to virtual reality is also apt to raise a host of philosophical concerns and worries. Perhaps one of the more pressing concerns comes from a consideration of what is lost during the process of biological death. The loss of the biological body is particularly worrisome, since there are reasons to think that the body is a crucial component of the human cognitive system, yielding a range of opportunities for intelligent action (Clark 2008) and mediating our emotional responses to both actual and counterfactual states-of-affairs (Damasio 1996). Such insights are not lost on those who work from within the PP camp. Seth (2013), for example, suggests that the processing of interoceptive information deriving from the non-neural bodily environment is relevant to some aspects of conscious experience, with a variety of "subjective feeling states (emotions)…arising from actively-inferred generative (predictive) models of the causes of interoceptive afferents" (Seth 2013, p. 565). Similarly, Hohwy and Michael (2017) present an intricate and intriguing account of the role of the biological body in giving rise to a sense of self.

Relative to these claims, it is far from clear that the attempt to replicate the generative capacities of the biological brain will be enough for the digital afterlife, especially if what we want to achieve is a state-of-affairs in which a given individual is resurrected as a sentient being, capable of enjoying (and enduring) the rich panoply of emotional states and conscious experiences that characterized their biological life. It is in this sense, perhaps, that the present account of digital immortality may be seen to be inadequate, focusing, as it does, on the biological brain at the expense of a larger material fabric that includes the individual's biological body and certain aspects of their local extra-organismic environment (see Clark 2008). The critic will no doubt want to highlight the indispensable role of the biological body in realizing certain aspects of the human cognitive economy, with a disembodied form of intelligence perhaps counting as no form of intelligence at all. They will also be inclined to view the afterlife options listed above as failing to address this problem. Talk of insubstantial hologrammatic ghosts and virtual world simulations are unlikely to do justice, they will say, to the role the biological body plays in shaping (and perhaps even realizing) the complex array of experientially-potent states that to a large extent make our lives worth preserving in the first place. From this perspective, perhaps the very best we could hope for would be some form of experientially-diminished afterlife—one in which our continued existence (if we care to call it that) comes at the expense of an ability to experience emotional states or to even have a sense of oneself as an entity that continues to exist. Perhaps a hologrammatic ghost, for example, is a prime candidate for a virtual version of what is dubbed the Cotard delusion—a psychiatric disorder in which the affected individual holds the delusional belief that they are dead or do not exist (Young and Leafhead 1996). This seems particularly likely in the wake of recent analyses of the Cotard delusion, which link the delusion to anomalies in bodily experience (Gerrans 2015) or

open the door to a more concrete form of digital afterlife, one in which the biological body is substituted with a synthetic, but no less substantial, corporeal presence.

aberrations in the processing of interoceptive information (Seth 2013). Of course, a delusion is only a delusion if the convictions of the relevant individual do not align themselves with reality. In this sense, it is doubtful that there are any genuine cases of the Cotard delusion in the afterlife. Believing you are dead in the afterlife is not delusional; what is delusional is to believe that you are alive when you are, in fact, dead. (No one said that the discipline of thanato-psychiatry would be straightforward!)

There is no doubt much here that is contentious, and I will not have the space to cover (let alone resolve) all the issues that are likely to animate future discussions in this area. It is worth noting, however, that nothing in the PP approach to digital immortality seeks to deny the importance of the body in mediating our cognitive engagements with the world, shaping our emotional responses, or, indeed, realizing aspects of conscious experience. In this sense, the biological body remains an important aspect of the human cognitive economy and a relevant target of generative models—hence the emphasis on tracking body-related information in Sect. 9.5. What is perhaps more problematic is the extent to which the various forms of afterlife I have described—holograms, virtual characters, and so forth—are properly characterized as lacking a body. Here I suggest that it pays to make a distinction between what Wheeler (2013) dubs *implementational materiality* (which involves a commitment to the idea that the body is no more than a material realizer of functionally specified cognitive roles) and *vital materiality* (according to which the body makes a non-substitutable contribution to cognitive states and processes). It should be clear that given the choice between these two options, it is only the commitment to vital materiality that poses any real threat to the prospect of virtual forms of embodiment. From a functional standpoint, therefore, I suggest that there is no real reason to regard a holographic entity or the inhabitant of a purely virtual world as congenitally condemned to a disembodied existence. Providing the functional contributions of the biological body can be replicated in virtual form, a hologram, I submit, counts as just as much an embodied entity as does an agent that has a more substantive physical presence.

9.7 Conclusion

The present chapter outlines an approach to digital immortality that is rooted in recent advances in theoretical neuroscience and machine learning. In line with other approaches to digital immortality, the present proposal highlights the importance of collecting and storing data about an individual, with a view to using that data for the purpose of digital resurrection. The difference between the present proposal and other accounts relates to the kind of data that is acquired, the way in which the data is analyzed, and the kind of computational substructure that is deemed relevant to the digital immortality agenda. The claim is that some hierarchically-organized PP system—some variant of today's deep learning systems—could engage in a form of lifelong learning, attempting to build generative models that tackle the same sort of

predictive challenges as those confronting the biological brain. Such models, it is suggested, will—by dint of the attempt to minimize prediction error—resemble those acquired by the biological brain as it attempts to secure a predictive grip on the sensorium. Inasmuch as we see these synthetic generative models as capturing the essential elements of who and what we are—models whose generative capabilities reflect our own biological capacity to render our realities, recall our pasts, and create our futures—then they may provide the means by which some aspect of ourselves is able to persist long after the biological body has withered away. The claim, in short, is that a hierarchically-organized predictive processing machine may serve as a vehicle that sustains our dreams as we inexorably succumb to the "sleep of death."

And what of poor Hamlet and his post-existential woes? Hamlet wonders whether it is better for him to die than to face up to his earthly troubles. But he worries that his death will be occasioned by dreams that merely serve to prolong his suffering. It is at this point, of course, that issues of technical feasibility come face-to-face with a host of more normative concerns. Just because digital immortality is possible (if, indeed, it is possible), does this mean that we should seek to make it actual? As a species we have done our best to preserve human life, and we have, I suppose, become somewhat good at it (even if many other biological species have had to pay the price). But have we done enough to ensure that the world in which we live is one that is worth living, as opposed to one that is worth leaving? Perhaps, then, the issue that lies at the heart of debates about digital immortality is not so much the technical obstacles that lie on the road ahead, as whether the route to digital immortality is one that is itself worth pursuing. Should we rage, rage against the dying of the light and resist the rule of the second law? Or should we accept that all dreams must end in a darkened room? To dream? To die? To be, or not to be? Ay, there's the rub.

Acknowledgements I would like to thank two anonymous referees for their helpful comments on an earlier draft of this material. This work is supported under SOCIAM: The Theory and Practice of Social Machines. The SOCIAM Project is funded by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/J017728/1 and comprises the Universities of Southampton, Oxford, and Edinburgh. Additional support was provided by the UK EPSRC as part of the PETRAS National Centre of Excellence for IoT Systems Cybersecurity under Grant Number EP/S035362/1.

References

Banerji, M., Lahav, O., Lintott, C. J., Abdalla, F. B., Schawinski, K., Bamford, S. P., Andreescu, D., Murray, P., Raddick, M. J., & Slosar, A. (2014). Galaxy Zoo: Reproducing galaxy morphologies via machine learning. *Monthly Notices of the Royal Astronomical Society, 406*(1), 342–353.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning, 2*(1), 1–127.

Betzel, R. F., Avena-Koenigsberger, A., Goñi, J., He, Y., de Reus, M. A., Griffa, A., Vértes, P. E., Mišic, B., Thiran, J. P., Hagmann, P., van den Heuvel, M., Zuo, X. N., Bullmore, E. T., & Sporns, O. (2016). Generative models of the human connectome. *NeuroImage, 124*, 1054–1064.

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. New York: Oxford University Press.

Clark, A. (2013a). Expecting the world: Perception, prediction, and the origins of human knowledge. *The Journal of Philosophy, 110*(9), 469–496.

Clark, A. (2013b). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–253.

Clark, A. (2016). *Surfing uncertainty: Prediction, action and the embodied mind*. New York: Oxford University Press.

Committee on the Analysis of Massive Data. (2013). *Frontiers in massive data analysis*. Washington, DC: The National Academies Press.

Damasio, A. R. (1996). *Descartes' error: Emotion, reason and the human brain*. London: Papermac.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM, 56*(12), 64–73.

Dieleman, S., Willett, K. W., & Dambre, J. (2015). Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society, 450*(2), 1441–1459.

Faisal, A. A., Selen, L. P., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews Neuroscience, 9*(4), 292–303.

Fawcett, T. (2015). Mining the quantified self: Personal knowledge discovery as a challenge for data science. *Big Data, 3*(4), 249–266.

Foroughi, J., Mitew, T., Ogunbona, P., Raad, R., & Safaei, F. (2016). Smart fabrics and networked clothing: Recent developments in CNT-based fibers and their continual refinement. *IEEE Consumer Electronics Magazine, 5*(4), 105–111.

Friston, K. (2002). Beyond phrenology: What can neuroimaging tell us about distributed circuitry? *Annual Review of Neuroscience, 25*(1), 221–250.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138.

Friston, K., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology, 3*(130), 1–7.

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: The brain as a phantastic organ. *The Lancet Psychiatry, 1*(2), 148–158.

Gerrans, P. (2015). All the self we need. In T. K. Metzinger & J. M. Windt (Eds.), *Open MIND: Philosophy and the mind sciences in the 21st century* (pp. 1–19). Frankfurt am Main: MIND Group.

Goertzel, B., & Ikle', M. (2012). Special issue on mind uploading: Introduction. *International Journal of Machine Consciousness, 4*(1), 1–3.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger (Eds.), *Advances in neural information processing systems,* Montreal, Canada (Vol. 27, pp. 2672–2680).

Hayworth, K. J. (2012). Electron imaging technology for whole brain neural circuit mapping. *International Journal of Machine Consciousness, 4*(1), 87–108.

Hinton, G. E. (2007a). Learning multiple layers of representation. *Trends in Cognitive Sciences, 11*(10), 428–434.

Hinton, G. E. (2007b). To recognize shapes, first learn to generate images. In P. Cisek, T. Drew, & J. Kalaska (Eds.), *Computational neuroscience: Theoretical insights into brain function* (Vol. 165, pp. 535–547). Amsterdam: Elsevier.

Hinton, G. E. (2010). Learning to represent visual input. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 365*(1537), 177–184.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313*(5786), 504–507.

Hohwy, J., & Michael, J. (2017). Why should any body have a self? In F. de Vignemont & A. J. T. Alsmith (Eds.), *The subject's matter: Self-consciousness and the body* (pp. 363–391). Cambridge, MA: MIT Press.

Hoogendoorn, M., & Funk, B. (2018). *Machine learning for the quantified self*. Cham: Springer.

Horn, B. K. P. (1977). Understanding image intensities. *Artificial Intelligence, 8*(2), 201–231.

Kugler, L. (2015). Touching the virtual. *Communications of the ACM, 58*(8), 16–18.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M. J., Nichol, R. C., Szalay, A., Andreescu, D., Murray, P., & van den Berg, J. (2008). Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society, 389*(3), 1179–1189.

Lupton, D. (2013). Understanding the human machine. *IEEE Technology and Society Magazine, 32*(4), 25–30.

Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., & Svetnik, V. (2015). Deep neural nets as a method for quantitative structure – activity relationships. *Journal of Chemical Information and Modeling, 55*(2), 263–274.

Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity*. Cambridge, MA: MIT Press.

Mitteroecker, P., & Huttegger, S. M. (2009). The concept of morphospaces in evolutionary and developmental biology: Mathematics and metaphors. *Biological Theory, 4*(1), 54–67.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data, 2*(1), 1–21.

Phan, N., Dou, D., Wang, H., Kil, D., & Piniewski, B. (2017). Ontology-based deep learning for human behavior prediction in health social networks. *Information Sciences, 384*, 298–313.

Ravanbakhsh, S., Lanusse, F., Mandelbaum, R., Schneider, J. G., & Poczos, B. (2017). Enabling dark energy science with deep generative models of galaxy images. In S. Singh & S. Markovitch (Eds.), *Thirty-first AAAI conference on artificial intelligence* (pp. 1488–1494). San Francisco: AAAI Press.

Revonsuo, A. (1995). Consciousness, dreams and virtual realities. *Philosophical Psychology, 8*(1), 35–58.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences, 17*(11), 565–573.

Seung, H. S. (2012). *Connectome: How the brain's wiring makes us who we are*. New York: Houghton Mifflin Harcourt Publishing Company.

Smalley, D., Poon, T.-C., Gao, H., Kvavle, J., & Qaderi, K. (2018). Volumetric displays: Turning 3-D inside-out. *Optics and Photonics News, 29*(6), 26–33.

Someya, T., Sekitani, T., Iba, S., Kato, Y., Kawaguchi, H., & Sakurai, T. (2004). A large-area, flexible pressure sensor matrix with organic field-effect transistors for artificial skin applications. *Proceedings of the National Academy of Sciences, 101*(27), 9966–9970.

Sporns, O., Tononi, G., & Kötter, R. (2005). The human connectome: A structural description of the human brain. *PLoS Computational Biology, 1*(4), e42.

Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data, 1*(2), 85–99.

van Gerven, M. A., de Lange, F. P., & Heskes, T. (2010). Neural decoding with hierarchical generative models. *Neural Computation, 22*(12), 3127–3142.

Wang, Y., Wang, L., Yang, T., Li, X., Zang, X., Zhu, M., Wang, K., Wu, D., & Zhu, H. (2014). Wearable and highly sensitive graphene strain sensors for human motion monitoring. *Advanced Functional Materials, 24*(29), 4666–4670.

Wheeler, M. (2013). What matters: Real bodies and virtual worlds. In I. Harvey, A. Cavoukian, G. Tomko, D. Borrett, H. Kwan, & D. Hatzinakos (Eds.), *SmartData: Privacy meets evolutionary robotics* (pp. 69–80). New York: Springer.

Yokota, T., Zalar, P., Kaltenbrunner, M., Jinno, H., Matsuhisa, N., Kitanosako, H., Tachibana, Y., Yukita, W., Koizumi, M., & Someya, T. (2016). Ultraflexible organic photonic skin. *Science Advances, 2*(4), e1501856.

Young, A. W., & Leafhead, K. M. (1996). Betwixt life and death: Case studies of the Cotard delusion. In P. W. Halligan & J. C. Marshall (Eds.), *Method in madness: Case studies in cognitive neuropsychiatry* (pp. 147–171). New York: Psychology Press Ltd.

Paul Smart is a Researcher at the New University of Lisbon, Portugal; a senior research fellow at the University of Southampton, UK; and a freelance cognitive science consultant, specializing in machine learning and virtual reality technologies. His research interests lie at the intersection of a range of disciplines, including philosophy, cognitive science, and computer science. He is particularly interested in the cognitive scientific significance of emerging digital technologies, such as the Internet and the Web. Paul's work has appeared in a number of journals, including *Minds and Machines*, *Phenomenology and the Cognitive Sciences*, *Synthese*, and *Cognitive Systems Research*. He is a co-author of *Minds Online: The Interface between Web Science, Cognitive Science and the Philosophy of Mind* and a co-editor of *Blade Runner 2049: A Philosophical Exploration*.

Part III
The Epistemology, Ethics and Deep History of the Extended Mind

Chapter 10
What Is It Like to Be a Drone Operator? Or, Remotely Extended Minds in War

Marek Vanžura

10.1 Introduction

They saw a flash in the sky–enemy aircraft. The hands of a pilot and a sensor operator started to sweat; a target they were nervously awaiting finally appeared in their sight. After a few seconds, their radar locked on to the enemy target, and the missile was armed and ready to be fired. But suddenly, an incoming missile pierced through the sky capturing the eyes of the crew. The flash that they previously saw was very likely its original moment of launch; the enemy pilot had spotted them first. There was nothing to do. In a fraction of a second, they were hit. Their aircraft burst into a wild fireball, beginning its ungraceful fall towards the ground below.

The pilot and sensor operator looked at each other. With visible relief on their faces, they were honestly happy that they did not sit in that plane. They could not. It was a drone, a remotely piloted aircraft without a human being on board.

Perhaps in this way we can reconstruct what happened in the skies of Iraq the day before Christmas Eve in 2002. For the very first time in history, a dogfight between a piloted airplane and an unmanned airplane took place. That time, the piloted one came out as the winner (see Kreuzer 2016). Despite this unfavorable outcome for the remotely piloted vehicle, this clash signified a monumental shift, namely, the rise of unmanned technology in twenty-first century warfare. During that above-mentioned battle, a drone, as those unmanned airplanes are commonly referred to, proved itself to be an invaluable protective measure of human lives–or did it?

Of the numerous technologies that shape our lives in the twenty-first century, drones are among the most controversial. What is usually studied is their impact on society or the communities living under their presence. Significantly less attention is paid to those on the opposite side of the technology: the human beings sitting behind the controls, the so-called drone operators. In the rare events that their work

M. Vanžura (✉)
Autonomous Driving Department, Transport Research Centre, Brno, Czech Republic
e-mail: marek.vanzura@cdv.cz

R. W. Clowes et al. (eds.), *The Mind-Technology Problem*, Studies in Brain and Mind 18, https://doi.org/10.1007/978-3-030-72644-7_10

does receive attention, it is often accompanied by disparaging and rude remarks such as "joystick jockeys" (Sparrow 2009, p. 25), "cubicle warriors" (Royakkers and Van Est 2010, p. 290), and "Playstation mentality" (Crilly 2010). These comments, however, only highlight the most obvious part of drone operators' job: remote operation. Among military personnel, remoteness is typically also highlighted but as the most advantageous aspect of drone operations. As illustrated by the opening story, drones can save the lives of their remote pilot, but the broader story is not that simple. There are also unacknowledged costs involved. As recent studies and various testimonies of drone operators reveal, drone operators pay for this imagined safety with their mental health.

In this chapter, we explore the world of drone operators, and inspired by a classic paper from Philosophy of Mind–Thomas Nagel's paper "What Is It Like to Be a Bat?" (Nagel 1974)–we explore a similar question: What is it like to be a drone operator? I will consider the work of drone operators, specifically in relation to their frequently experienced post-traumatic stress disorder, which is both striking and disturbing. I propose a hypothesis that is meant to describe factors and causes that affect drone operators. Based on this hypothesis, it should then be possible to design proper empirical research which might lead to a better understanding of this new profession and related phenomena. Philosophy is a perfect tool for investigating all of these possible directions beforehand. I argue that remotely operated vehicles become an extension of their operators' minds, cognitive faculties, and very likely also their emotions. In other words, a technology of unmanned, remotely piloted vehicles is another case where a mind extends beyond the skin and skull of a human being; the minds of drone operators are remotely extended into drones they control and operate. This means, and elucidates, that while unmanned remotely controlled technology has great potential to protect physical aspects of life, there is still a great risk of psychological suffering here.

In the first section, I give a brief introduction into the world of remotely piloted aircrafts, with special attention to their relations with human operators. The second section goes more into depth about the work of a drone operator, including data on post-traumatic stress disorder among drone operators. In the third section, I present my own explanation of the cause of this phenomenon: operators are prone to mental health issues previously connected only with physically deployed soldiers for the reason that drones quite literally extend the cognitive functions and minds of their operators. Therefore, drone operators are yet another instance of the extended mind concept. In the fourth section, I elaborate on one specific factor that seems to be highly important in affecting minds of drone operators, viz. the role of a causal connection, or more precisely, a role of perceived causal connection. I propose a scenario, or thought experiment, aimed to uncover how to deal with the mental problems of drone operators. Nevertheless, I also examine ethical arguments that challenge potential efforts to make the working environment of drone operators truly risk-free. Finally, the fifth section anticipates potential objections against my proposed explanation and offers my replies to them. This final part summarizes the entire chapter and answers the central question: What is it like to be a drone operator?

10.2 Technology of Remotely Piloted Aircraft

Remotely piloted aircrafts per se are not new. We can trace their roots back to the 1930s, when British military was looking for aerial target that could be used for practicing anti-aircraft gunnery and dogfights. The solution was the airplane called Queen Bee (see Draper 2011). Radio control back then was rather crude, but the idea proved to be useful and promising. During World War II, the first experiments with drones being flown through television camera emerged. This was the TDR-1 project, which had its war debut in 1944 in the Pacific theatre against the Japanese (see Hall 1984). During the following decades, remote control was used mainly for flying aerial targets. Throughout the 1960s and 1970s, during the conflict in Vietnam, unmanned technology took a slightly different path. The pilotless aircraft called Firebee and Lightning Bug made more than 3000 raids over Vietnam with the intention to collect information, primarily take aerial photos, for intelligence purposes (see Wagner 1982). During that time, these planes were not controlled remotely, but pre-programmed instead. This solution enabled flights to go further into enemy territory, because it did not have to rely on unstable and still not very reliable remote connections. Firebees were in a sense autonomous–for a period of time at least (cf. Mindell 2015).

Ongoing technological progress, however, brought important abilities to overcome obstacles that had limited unmanned vehicles so far. At the end of the twentieth century, there were numerous technological advancements, such as satellites and increased computing power in reasonably small devices, that made possible the birth of the first version of the later, infamous Predator drones (see Whittle 2014). This was a turning point in unmanned technologies because they were now able to send video footage in real time to the commanding headquarter. This really was a revolution in the field of drones.

The history of remotely piloted aircrafts is therefore long, but modern drones are, nevertheless, different from those that preceded them. Unlike their predecessors, modern drones bring immersion. Thanks to their cameras with high resolution, optoelectronics with incredibly long focal lengths, stabilized sensor pods, precise software that allows to stare at one location or follow moving objects, and the ability to loiter over an area for 24 h a day–current drones bring to remote operators opportunities that allow them to feel like an "eye of God" (Chamayou 2015, p. 37). Operators feel they are present in the place they view from above. Yet, in reality, they look from afar, often through infrared cameras that give them unusual visual stimuli coupled with an aerial perspective.

All of the aforementioned working conditions would make drone operators just passive observers, almost like some kind of voyeurs–if we ignored their drones' interactive equipment. Given this additional facility, drone operators are able to affect the world they perceive on their screens. Their drones carry weapons, such as laser-guided missiles, that reap material consequences on what they see. Weaponry on drones is perhaps the most controversial aspect of current drones, although it is not all that new. Armed drones were, in fact, tested for the first time in the 1940s and

then again in the 1960s, but it was just a one-time episode with no direct continuation, mainly due to the underdeveloped state of technology at the time (see Wagner 1982). Thus, it is the twenty-first century that ultimately brought lethal drones on the scene.

10.3 Post-traumatic Stress Disorder and Drone Operators

A New York Times headline reads, "Drone Pilots Are Found to Get Stress Disorders Much as Those in Combat Do."[1] "Post-Traumatic Stress Disorder Is Higher in Drone Operators," reads another headline in The Telegraph.[2] "The Warfare May Be Remote But The Trauma Is Real," heads a National Public Radio story.[3] These headlines, published in the last decade, are just three samples of newspaper articles that are devoted to the impact of remote combat on drone operators. The theme of mental health of drone operators obviously produces a lot of interest among the general public. Although the popular media tends to exaggerate real numbers, it is still quite a disturbing situation.

Before we examine in more detail findings on the actual presence of post-traumatic stress disorder among drone operators, it is necessary to better understand the work of a drone operator.

The most famous drones or unmanned aerial vehicles are the American aircraft called MQ-1 Predator and MQ-9 Reaper. Those are armed and therefore potentially lethal machines which are easy to recognize due to their bulbous noses that cover antenna for satellite communication. The usual crew remotely manning these drones consists of a pilot responsible for flying the aircraft and a sensor operator whose job is to operate a stabilized sensor pod with optoelectronics (see Dougherty 2015). For simplification, I refer to both crew positions–pilot and sensor operator–as drone operators in this text. When a missile attack is requested, the sensor operator designates the target and illuminates it with a laser, whereas the pilot launches the missile. An attack is thus a cooperative effort. Along with the drone operators, there is also an analyst who shares the trailer space, but this person is not as directly involved in operations as the pilot and the sensor operator; therefore, I will not consider this person as a drone operator.

[1] Dao, J. (2013) Drone Pilots Are Found to Get Stress Disorders Much as Those in Combat Do. The New York Times, accessed September 19, 2017, http://www.nytimes.com/2013/02/23/us/drone-pilots-found-to-get-stress-disorders-much-as-those-in-combat-do.html

[2] Hawkes, R. (2015) Post-Traumatic Stress Disorder is Higher in Drone Operators. The Telegraph, accessed September 19, 2017, http://www.telegraph.co.uk/culture/hay-festival/11639746/Post-traumatic-stress-disorder-is-higher-in-drone-operators.html

[3] McCammon, S. (2017) The Warfare May be Remote But the Trauma is Real. National Public Radio, accessed September 19, 2017, http://www.npr.org/2017/04/24/525413427/for-drone-pilots-warfare-may-be-remote-but-the-trauma-is-real

Besides flying a drone, the job of drone operators consists of numerous hours of watching video feeds provided by sensors carried by the drone flying over the war zone. Operators may zoom in and out as they like, day in and day out, and they can change the position of their airplane to follow a person or a vehicle. They may orbit over an area of interest, and ultimately, they may engage a Hellfire missile, a laser-guided missile characterized by deadly precision (see McCurley and Maurer 2015). The continuous surveying also allows drone operators to become familiar with the life patterns of people they watch. Many of the drone operators confess that they even developed emotional attachments to the people they see on the ground (see Power 2013). And then, the order to strike arrives. That is to say, they have to conduct a lethal attack on the people they think they know.

There are also strikes called "double tap attacks" that are conducted shortly after the first attack, where the purpose is to kill those who came to help–the reasoning being that they are guilty as well. Thus, drone operators very often have to see the consequences of their actions, such as body removals. Such work gets very intimate and emotional even though it is conducted from thousands of kilometers away through computer monitors (see Martin and Sasser 2010). Hugh Gusterson calls this phenomenon a remote intimacy (see Gusterson 2015). The mass of video footage that drone operators see on their screens are sometimes called "war porn" (see Singer 2010). The footage is very immersive and also addictive. This immersive dimension of drone operators' work is of significant concern among psychologists and other medical personnel, particularly because of the presence of post-traumatic stress disorder and other psychological issues among drone operators.

The United States Air Force drone operators are a highly secretive and protected breed of military personnel, which makes it difficult to access proper and sufficient information about their mental conditions and health. The most prominent studies among drone operators were conducted by Chappelle and colleagues (see Chappelle et al. 2014a, b). These studies are possibly the only reliable source of data. And as the authors have admitted, their studies have certain limitations and have therefore obtained results that might not be as precise as one would hope, but they are the best we can obtain.

"PTSD [post-traumatic stress disorder] is a well-known psychological condition that may develop after exposure to a traumatic event (witness or experience events that lead to actual or threatened death, injury to others) where the individual experienced intense feelings of fear, helplessness, or horror" (Chappelle et al. 2014b, p. 64). As such, PTSD is mostly diagnosed in the victims of violent crimes, such as rape or sexual abuse, or among firefighters, police officers, and soldiers who faced imminent danger, where their lives were in danger of death.

Screening soldiers coming back from their deployment for symptoms of PTSD, either by the army, air force or navy, is common practice. On the other hand, drone operators who participate in war from afar are typically considered as stationed on the homefront and in a relatively peaceful environment, and therefore, presumably, not subjected to the necessary conditions that produce PTSD and other related disorders. Studies have shown, however, that even these "cubicle warriors," as drone

operators are sometimes deridingly called, are prone to suffering the same psychological problems as their deployed counterparts.

As mentioned above, Chappelle et al. have conducted a series of studies among drone operators. The most recent one dates back to 2014. A total number of 1084 USAF drone operators participated in the study, of which, as results later showed, a total of 4.3% displayed symptoms of PTSD (see Chappelle et al. 2014a). To put this in context, according to the meta-analysis conducted by Richardson et al., estimations of combat-related PTSD among soldiers returning from deployment varies from 4% to 17%, depending on the measurement methods (see Richardson et al. 2010).

Psychologists Armour and Ross (cf. Armour and Ross 2017) recently published a study summarizing every account of mental health on drone operators and intelligence analysts. Besides the work of Chapelle's and his colleagues, there are 14 other publications dedicated to this topic, although not all of them focus on PTSD. But even if we select just those concerned with PTSD, results of this comparison basically confirm the above-mentioned results.

In the case of US drone operators, levels of PTSD are not, as studies revealed, higher than in the case of pilots of manned aircrafts. Those levels tend to be the same or lower (see Wallace and Costello 2017). But it is still quite significant, given their physical safety. Popular media does tend to overestimate the significance of PTSD among drone operators, but this does not decrease its importance. These issues are evidently also attracting the attention of responsible personnel within the military. With the prospect of increasing numbers of drones, and therefore also of drone operators, it is very likely that there also will be an increasing number of cases of PTSD among them.

On the face of it, this is a paradoxical situation. Uniquely among participants in war, drone operators may for the first time in history participate in war in a physically safe environment, without the threat of any direct, violent consequences to their lives. Yet at the same time, they suffer psychologically like their less fortunate counterparts deployed in the middle of war zones, subjected to the physical horrors that entails. The obvious question arises: why is there a seemingly unsafe mind in a safe body? This can appear puzzling, especially if someone sees drone operators as mere videogame players. Such puzzles naturally attract the attention of philosophers. In this case, I think that philosophy can shed some light on the mystery.

10.4 Extended Mind and Drone Operators

I propose to consider this case as another manifestation of the extended mind concept. The extended mind theory basically says that some of our mental processes are realized by mutual interaction between our biological tissue and external artifacts (see Clark and Chalmers 1998). Although someone might find this theory controversial, I accept it as a plausible and sufficiently robust theory to build upon. Thus,

I do not use the situation of drone operators as proof that the extended mind theory is true. I presuppose it.

Moreover, the extended mind theory allows us, in contrast to other approaches and frameworks, to treat this situation within the current line of discussion without the necessity of reformulating already existing categories and definitions. In other words, we may explain the phenomenon of remote stress and trauma without the need to change psychiatric metrics, for example.

The authors of the extended mind theory, philosophers Andy Clark and David Chalmers, present a framework that allows us to understand under what circumstances we can understand some cognitive processes as extended. In short, when processes in the world have the same functions or outputs as processes usually attributed to cognition that takes place in the head, then those external processes might properly be called cognitive processes (see Clark and Chalmers 1998); this is the so-called "first-wave" of extended cognition (see Kirchhoff 2011, p. 289). But this does not mean that external processes must be exactly the same as internal processes. In fact, the point is that those external processes are usually very different from those happening within our biological boundaries. What is important, is the function. Richard Menary calls this cognitive integration (see Menary 2008); this is part of the "second-wave" of extended cognition (see Kirchhoff 2011, p. 290).

For some external process to be a part of someone's cognition or mind, there must be a causal coupling between them. When a human being is working with some external artifact that, in turn, affects him or her back (like a feedback loop), a coupling occurs. "In these cases, the human organism is linked with an external entity in a two-way interaction, creating a coupled cognitive system that can be seen as a cognitive system in its own right." (Clark and Chalmers 1998, p. 8).

Clark and Chalmers present this notion of coupled cognitive systems in a more systematic way as follows:

1. All the components in the system must play an active causal role: each part must causally affect other parts – that is, the internal part affects the external and vice versa.
2. All the components jointly govern the behavior in the same sort of way that cognition usually does.
3. If we remove the external component, the system's behavioral competence will drop, just as it would if we removed part of its brain (see Clark and Chalmers 1998, pp. 8–9).

This influence of components on each other must be symmetrical. In other words, internal and external parts causally influence each other mutually, not just exclusively one way or other. To better illustrate the importance of mutual influence, Clark introduced the concept of continuous reciprocal causation, which "occurs when some system S is both continuously affecting and simultaneously being affected by activity in some other system O." (Clark 2008, p. 24).

Notably, common sense typically attributes to internal processes significantly different properties than to external processes. Thus, there are perceived implicit characteristics of internal functions that separate them from external ones. Also,

thinking about extending cognitive and mental resources into external artifacts can potentially make extension too expansive. To avoid such mistakes, prejudices, or fallacies, Clark and Chalmers have developed conditions that are considered necessary for something to be called the extension of mind and/or cognition.

Those conditions are:

1. An external component must be a constant part of someone's life (reliability condition).
2. The information contained or provided by the external component is available without difficulty (availability condition).
3. The user automatically endorses the information provided by the external component (truthfulness condition). (See Clark and Chalmers 1998, p. 17)

This sums up the extended mind theory sufficiently for our present purposes. Let's now take a look at the condition of drone operators and how it coheres with the just reviewed theoretical background offered by the concept of the extended mind.

The nature of the drone control interface is of primary importance in evidencing casual coupling. Notably, the interface's design aims to create two-way reciprocal interaction; ironically, its earliest design iterations were often criticized for not achieving this goal sufficiently (cf. Asaro 2013). Nevertheless, it aims to be a tool that allows for a coupled cognitive system as far as possible.

Drone operators' sight becomes the drone's sight. Drones react to commands to change the direction of the flight, the direction of the camera or the magnification of the optoelectronical systems, etc. Through drones, drone operators manipulate the physical world, although, manipulation is mostly limited to mere destruction. Subsequently, the drones cameras and sensors offer back to the operators information that allows operators to assess and evaluate all the damage they caused and determine the next course of action. This assessment, moreover, is also complimented by military personnel on the ground who might collect artifacts for intelligence purposes, offering drone operators another channel to receive information about their drones, their surroundings, and the interaction between the two.

This all offers sufficient background to create new agent-world cognitive circuits. Each part of the system, the drone and its operator, is causally active. We can see that the relation between drone operators and drones, through the interface, is truly of a two-way nature, where both parts mutually affect each other. The sensing achieved by sensors on a drone becomes part of the larger cognitive process, which involves cognitive processes of the human being sitting behind the controls. Unsurprisingly, if we remove the external part (in this case, the drone) the behavior would change drastically. This is particularly evident because of the physical remoteness between the drone operator and the drone.

A drone is really not a mere mediation device, but an active cognitive component of a whole integrated system. It has numerous cognitive roles. Drones harbor external memory. Those technologies store data and information a human operator relies upon and acts upon. Without such information he or she would not be able to conduct a mission. Another cognitive role drones usually play is that of perception. They are equipped with cameras that allow drone operators to see. This might be

argued as being only a mediation, but that is not true if we consider infrared cameras. To see in infrared light spectrum is something humans cannot naturally do, but operators, thanks to drones' camera technology, are able to conduct missions even in the dark of night. If we remove a drone from this system, it affects the overall performance of cognition in the same ways that removing part of a brain would.

As I will now argue, a drone operator and his or her drone meets all the conditions for extended cognitive systems. First of all, consider the reliability condition. A trailer with the control interface for remote control over a drone is a constant part of the drone operator's life: it is a necessary component in the daily working conditions of drone operators. In the agency that they exercise at work, drone operators fully rely on their equipment, and given the current state of the technology, they can very reliably do so.

Secondly, there is the availability condition. This condition is fulfilled completely. Every bit of information drone operators need or require for their work is easily obtainable via the drone sensors and instantly displayed on their monitors. Operators have everything at hand, and the control and integration of it is so automatic that some have deemed it "transparent equipment" (Clark 2007, p. 267). Drone operators work through their controllers, not with them. It becomes so natural for operators to use them, that they do not think about it anymore. It is even possible to say that they use those controls the same way as they use their own brains.

The third and final truthfulness condition is satisfied as well. In the case of drone operators, they endorse the information that is presented to them, or to put it another way, they trust it as being both truthful and a precise, reliable representation of reality. They trust the information presented on their screens. In fact, there is no need or reason to doubt it. When they feel uncertain, they may fly a drone to a different position to get a better perspective and watch from a different angle. Moreover, the entire military complex authoritatively supports and reinforces the belief and feeling that everything they do is serious and that the scenes presented on those monitors are true images and representations of a real world, albeit somewhere very far away.

All of this, as I propose, should persuade us to see drone operators as part of coupled cognitive systems created together with their drones that function as extensions of their cognitive processes and minds. That is, all three main conditions are satisfied. Thus, we can conclude that drone operators' minds are extended out to their drones, via the operating systems within their controlling equipment in the trailers they work in. To fully explain why mental issues arise among drone operators, however, we also need to consider complementary factors in this system.

Another key factor to consider that is tightly connected to the extended nature of drone operators' mind is the phenomenon called a "place lag," a term coined by Mark Vanhoenacker (see Vanhoenacker 2015, p. 23). Much like the well-known phenomenon of jet lag that describes how a human body can be fooled by quick changes in time zones because of modern aviation that allows us to travel on long haul flights across vast distances relatively fast (see Lee 2017), place lag signifies a confusion caused by fast changes in cultural and other environments. Also experienced mainly by frequent flyers on long haul flights and airline pilots, as is

Vanhoenacker himself, a place lag is initiated by moving fast around the globe from one cultural sphere to another without gradual acclimatization. For example, someone can experience a place lag when he or she travels from a highly technological city such as Tokyo to places in central Africa. Differences in perceived cultures and other related factors are so huge that it can cause discomfort similar to that of jet lag.

Drone operators very likely experience place lag as well. That is so because drone operators who fly drones during their missions over war zones are stationed on U.S. soil. Usually based at the Creech Air Force Base in Nevada (see Cullen 2011), drone operators are surrounded by the conditions of American culture and life, accompanied by a relatively peaceful environment. Only when after they step into their trailers equipped with high-tech gadgets to fly drones do they step into a war zone. There is quite literally just a matter of a few steps that takes them from one cultural environment to another. The other environment is radically different, which is the perfect condition to induce a place lag. Doing this on a regular basis, it is very likely that such regular movements in and out of perceptions of present reality create a severe psychological dissonance.

Worth noting here is the contrasting case of Israeli drone operators. A recent study suggests that these operators are not likely to suffer from PTSD (see Gal et al. 2016). There are many possible explanations. It could be argued that the conditions for place lag to take place are not sufficiently occurring as their drone-extended minds are not working in a significantly different environment. The drones they operate are not deployed on the other side of the globe in a condition dissimilar to operators' larger cultural milieu. Drones operated by Israeli Defense Forces (IDF) fly over the Israeli territory or nearby, which means that operators do not experience sights different from their everyday lives. It is also worth considering, as it is often highlighted, that the State of Israel is in a permanent condition of war with its neighbors. This also necessitates that the majority of Israelis has to attend mandatory military training and service, which supports a sense of possible threat at any time, whether operating drones or not. Such a situation is very different from the situation of the American drone operator who flies drones over Afghanistan, for example, from the airbase in Nevada where they return back home daily in a completely peaceful country after a day's shift is complete.

This is, however, just one of many explanations. There are surely many others, for example, a notion of emotional attachment. Israeli drone operators might consider their job much more personal than American operators because of their involvement in war that is happening at their doorstep. At this point, we cannot rule out one option and confirm the other. Furthermore, complex phenomena are almost never a result of just one cause. Thus, it is much more reasonable to expect that there are many factors coalescing.

To summarize my proposal, I argue that the minds of drone operators literally extend beyond their physical bodies. This fact plays a significant role in the psychological states and responses of drone operators to their work. Together with place lag, this is likely a key cause of mental disorders, such as post-traumatic stress disorder, occurring among some drone operators. Also, as Chappelle et al. argue, another factor is operational stress, like working in shifts, insufficient numbers of

personnel and so on (see Chappelle et al. 2014a). When compiled together, we begin to get a satisfactory explanation of the most likely causes of mental disorders among drone operators and clarify the intricacies of what it means to be a drone operator. Furthermore, adding in extended mind theory explains why the mind is in danger of being exposed to the same mental traumas of physically being in a war zone, even though the human body of the drone operator is not.

10.5 A Role of Causal Connection

A little more attention should be paid to the causal connection, or more precisely how this causal connection is perceived by drone operators as it seems to be a crucial point in further developing mental problems from this remote presence in war zones. I propose to consider the following thought experiment, which might be turned into a real empirical one–although, it may be morally troubling to conduct it. This experiment should also help us to discern at what moment there is not enough reciprocity to create a coupled cognitive system.

Let's consider the situations of three drone operators. The first one is basically a perfect example of current drone operators. She is completely responsible for every decision she makes, including weapons release. That means, she is the one to decorate or blame for killing intended or unintended targets. In other words, she is a clear part of a causal chain.

The second drone operator, however, is in a less responsible position. Her drone is highly automated, and sufficiently advanced to make decisions on its own, including those of target acquisitions and releasing weapons, which usually leads to the death of people. The drone operator in this case only supervises the drone, seeing if it works according to the rules of engagement, for instance. If not, this drone operator is able to intervene and stop the attack. In this situation, a drone operator is not directly included in a causal chain, but she might stop the attack – which may save lives or endanger lives if this intervention allows the enemy to conduct an attack on the operator's friendly forces.

Finally, the third operator is totally out of the loop. Her drone is fully autonomous and makes all decisions by itself without human supervision or intervention. It is basically an agent of its own. The human operator is reduced to the position of a viewer and removed out of a causal chain.

The question is: Will there be any significant difference between those three situations concerning effects on the operators' minds? If so, what differences are there? Basically, we could expect two main outcomes. First possible outcome: there would be no change in the operators' reactions. This is very unlikely. The perceived sense of responsibility and associated burdens are so different between the particular scenarios that such a result is almost impossible. Second possible outcome: there would be significant changes in behavior and mental states from one scenario to the next. It is reasonable to assume that the less a drone operator is involved in the machine's

decisions and actions, the less he or she is bothered with the consequences of the witnessed events.

In other words, relating this back to the extended mind theory-based proposal, the removal of the operator's agency from the causal chain would mean terminating the extension of his or her mind. Therefore, reciprocity and interactivity are necessary for extending the mind into a drone. Immersion is thus a function of interactivity. Mere witnessing is not enough. From a practical standpoint this means that engineers and designers should be able to modulate drone operators' reactions through the modification of the control interface. This could be done in many ways. As shown in the thought experiment above, it can be done by increasing automation until basically full autonomous, for instance. Another approach could be implementing software tweaks downgrading the quality of images, or increasing a game-like appearance, or disengaging the moral involvement of operators by pretending it was the computer system's fault that led to the bloody events, among many others. These are partly engineering challenges– but, there are also ethical challenges that we have to consider.

Considering the extended nature of drone operators' minds, modifying a response through interventions into components that play a constitutive role in the extension of the human mind might be morally troubling. As shown in a similar context by Neil Levy (cf. Levy 2007), this could be viewed the very same way as interventions into the operator's physical body. In other words, it might be considered as a personal assault (cf. Carter and Palermos 2016). However, it may not be negative if this change is done to help drone operators with their psychological problems. For designers of these interfaces and their users, it is, surely, preferable to eliminate all the negative effects such an interface can induce, but from a broader perspective, it is an open question whether it is really desirable to remove this (psychological) burden that is put on drone operators. Critics of drone warfare might argue that war is not a game, and if we remove all negative consequences that remotely conducted wars now have, then we may end up with drone operators who view their work as an amusing game. Therefore, it might be argued that maintaining some sobering effects on operators should be mandatory. Alas, they are still soldiers–the very notion of which assumes a certain personal risk.

The discussion of reducing or increasing impact/burden through modifying external parts of drone operators' minds probably cannot just center on the interface, but would also need to consider the drones as well. In other words, from the perspective of this author's proposal, threats to drones flying over war zones would be considered threats to their operators themselves. Saying that, someone might argue that drone operators physically face threats of war as well (cf. Carter and Palermos 2016). Admittedly, this proposition is very controversial.

10.6 Objections and Replies

Several possible objections against my proposal can be envisioned. Because the extended mind theory is still a rather controversial concept, the most obvious form of criticism would be against the extended mind theory itself (e. g. Adams and Aizawa 2010; Rupert 2009). However, despite the severity such general arguments may have, my focus here is on potential direct objections toward my specific proposal: objections against the application of the extended mind framework to the situation of drone operators.

Objection 1: The extended mind concept is redundant. We really do not need it in order to explain why drone operators suffer from mental health issues. In fact, the situation is much simpler. They suffer, because they kill people and have to watch it on their screens. Their minds do not extend; they are traumatized just because they see all those horrors.

Answer: This objection is very important and serious and requires adequate attention. It is tempting to accept such a simple explanation and refuse to accept a more complicated alternative. However, I claim that we cannot sufficiently grasp this situation without the emphasis on reciprocal causal connections that are involved in the extended mind framework and that are lacking in the simpler frameworks. Although the mere watching of traumatic events can in some cases generate an intense outcome such as post-traumatic disorder or other problems (see Pinchevski 2016), the interactivity and being in the loop seems to extend beyond this simple explanation in the case of drone operators. Not to mention, using the extended mind theory framework, it is easier to describe the nature of the drone operators' experiences in currently used psychological terms without the need to drastically reformulate definitions and notions of involved phenomena. I mean, thanks to the extended mind theory, we are able to describe the situation of drone operators in current terms, which would not be that easy if we would consider, for example, a nature of witnessing in a purely mediated situation. In the proposed simpler variant, we have to explain what it means to witness something if someone does it by mere watching a screen, which goes beyond what is traditionally understood as witnessing. My proposal, however, bridges this problem by accepting that a certain part of a remote drone operator is witnessing a situation directly. This expansion of a drone operator's mind into a drone then allows us, for instance, to apply standard notions of witnessing.

Even if someone experiences feelings of fear or horror during watching a movie or live coverage from some unfortunate part of the world, it is always possible to detach from it and tell yourself that it is not happening, or it is happening somewhere else, images on a screen are just a mediation of these events. On the other hand, drone operators cannot detach themselves the same way, because they repeatedly participate in the events, although remotely.

An opponent might at this point raise another objection that reads as follows. There is not a big difference between a drone operator and a soldier who launches a

cruise missile by pressing a button. They both work remotely and watch their outcome on screen, either in a ground control station as in the former case, or on television as in a latter case, but a soldier is not extended to that missile. Why should a drone operator be considered as extended to a drone? I would argue those situations are not on a par. That is because of the fact that drone operators are involved in a continual feedback loop with their drones. On the other hand, a soldier pushing a button to launch a missile, who then sees footage of what it did, is causally separated by a lack of continuous connection. In this case, the cause and effect are perceived as two discrete situations. The link between them is missing. It is harder for a soldier to feel responsible or guilty, because there is an amount of time he was not in control. Drone operators do not experience this fragmentation of events.

Also, the fact that they kill people remotely is not enough to justify that they feel engaged. In unison with drone critics we may say that drone operators just push buttons, which is relatively easy to do without the sense of participation. But through the extended mind theory's lens, we are able to grasp and explain it with much better precision. Drone operators feel involved due to the immersion they experience, which leads to the sense of "being there". Their mind is relocated: it is partly at the controls in a station and at the same time, it is partly hovering over the war zone in a drone. The resulting mental state is comprised of one part coming from a drone, that is an external technological artifact, and another part coming from a biological circuitry of the drone operator's human body and brain. Not to mention that the proposed alternative explanation does not give us any hint regarding how to deal with the situation and how to modulate outcomes and effects on people who do this job. I therefore believe that such a simple explanation is only a starting point from which we have to move to a more complex explanation. In this case to the extended mind theory.

In sum, although it is possible to consider the easier approach proposed by the objection alongside the approach of the extended mind theory proposed by this chapter, the latter offers more robust tool for further work.

Objection 2: The role of being a drone operator is constrained by time on many levels. On the larger scale, someone works as a drone operator for a limited period of time during his or her lifetime. But our minds accompany us for a whole life. In other words, drone operators bring their own mind into the job and when they retire, they still have their mind, which does not include drones anymore. So, their mind is not really extended, because it does not continue afterwards. On a shorter scale, the work of a drone operator is done in shifts. Therefore, your proposal says that their minds are extended when they arrive at their job and then reduced to their natural size when they leave the door of a ground control station. Also, different persons work with the same interface and drone. If your proposal is right, then their minds overlap. They are constituted by exactly the same components which they share. Needless to say, there is a chance of being shot down which would consequently lead to the loss of the operator's mind, at least the extended part.

Answer: First of all, as far as I am aware, no extended mind theory proponent claims that external components of coupled cognitive systems are necessarily part of this system for its entire lifetime. In fact, the advantage of this coupling is that it is realized only when it is beneficial. The main cognitive machinery is still in the head. But there are some situations or tasks that are better accomplished when someone creates a coupled cognitive system with some artifact in the external environment. That is the moment when coupling takes its place.

More concretely, a mind is not a solid object. It changes all the time, mainly because of the phenomenon called neural plasticity. When we were born, we had to learn how to speak. During our lifetime, we gain as well as lose many abilities of our cognitive apparatus and minds. Therefore, it is natural that a mind has some properties at one time and different properties at another. There is really nothing abnormal when minds of drone operators are periodically extended during their shifts and then shrink back when their working hours are over. The same applies to the larger scale. When they retire, they do not use this particular extension anymore, so their neural networks may rewire for some new task.

I do not see any problem with overlapping minds. Moreover, it is very likely a good thing. When we consider language, everyone who understands, say, English has a basic vocabulary that is shared across the English-speaking community. Although it is realized in separated heads, the underlying component of abstract words is shared. In the end, those shared constitutive components are necessary for mutual understanding. Similarly, when drone operators share their control interface and drones with other operators, it helps them to maintain a professional group identity which they share.

Perhaps the most severe objection among these is one that highlights a possibility of losing part of a mind when a drone gets shot down. This is a real problem, to which I can offer the following answer. A drone operator might be willing to lose a part of his or her mind in a case of enemy action, because there is only a slight chance that such situation will really occur. So, it is an acceptable risk. It is similar to those people who intoxicate themselves with alcohol, even if they know it harms their brains. Such people rationalize that this is tolerable because of related positive effects, such as a short-term improvement in the mood. In the same vein, drone operators might tolerate this because of perceived benefits, such as protection of ground troops. So, yes, my proposal involves the danger that drone operators lose parts of their minds if their drone is shot down, but there are ways in which this is acceptable.

Lastly, virtually every of these criticisms may apply to other technologies, such as mobile phones, that might be lost, stolen or lose reception. Yet, mobile phones are the picture-perfect instance of extended cognition. Hence, these objections do not present significant obstacle for the extended mind theory.

Objection 3: The studies on the occurrence of PTSD among drone operators show that they experience symptoms of this disorder the same level or even lower than their counterparts, such as pilots of manned aircraft. Let's thus accept some similarity. But what significantly differs, is the way these two groups experience the

war that leads to their mental suffering. Your proposal argues for immersion to be a key element in drone operators, but pilots inside a manned aircraft lack this kind of immersive experience created by perceived visual proximity. They sometimes do not even see the target, it might be just a GPS coordinate to them, whereas drone operators see the whole sequence of killing. In short, mechanisms involved in generating mental trauma in these groups are different.

Answer: That is a very good point. It seems like two different kinds of mechanisms, but I believe there is an underlying similarity. A pilot flying in a supersonic jet over a war zone is a potential subject of enemy fire. Thus, he or she is exposed to the risk of being killed. Whereas a drone pilot flying a drone remotely from the safe distance is seemingly not. It is immersion that invokes in a drone pilot the sense of being there, being on board of a drone and flying over enemy territory. A drone operator is–thanks to the cameras–so close to war that he or she might find himself or herself experiencing "being there." Although a pilot of a manned aircraft is physically present to the potential threat, a drone operators' presence is simulacra, since they are remote. But it is still sufficient to invoke a fear of being killed or harmed when things get intense. Support for this claim comes from an account of a drone operator for whom the experience was so real that when his drone was under enemy fire and started to fall down from the sky, he instinctively reached for a handle on his seat to eject himself out of this damaged aircraft (see Rogers and Hill 2014, p. 72). In other words, they both care for their lives, which is the underlying factor that leads to the eventual mental trauma. A pilot of a manned aircraft is concerned about his life because of his physical presence in a dangerous place, a drone pilot is concerned about his life because of his perceived presence in a dangerous place thanks to the immersion a drone brings. The process behind each of these cases is different, but their outcome is equal.

To put this into a wider perspective, this fact plays an important role when we consider a relationship between mental issues and a relatively long extension beyond their physical bodies. Such a long extension is necessary to fully understand what is hidden behind the problems drone operators experience. A shorter extension does not suffice for the above-mentioned concern for someone's life when it comes to critical situations, such as being shot down.

10.7 Conclusion

Interaction between the human brain and mind, and the twenty-first century technologies and artifacts are sometimes straightforward and, at other times, incredibly complicated. The technology of remotely operated military vehicles is a perfect example of the second category. It is a situation that challenges previously used frameworks and surpasses old boundaries– this is not necessarily a bad thing. It prompts us to better understand ourselves as well as our creations.

What the study of drone operators uncovers is the fact that effects of a technology on human beings are not often as simple as is generally accepted. Drones are promoted as unmanned and risk-free to their users. But it is in fact only partially true. They are unmanned in one sense, but not in another. There are still people who use, control, and maintain this technology. They are risk-free in one sense, but not in another. There are still operators who suffer from the same disorders as soldiers who were deployed in harm's way. To really understand our technologies is to address all of its faces and consequences. If we reformulate the famous question asked by Thomas Nagel, which has been resonating in the field of philosophy of mind for more than four decades, to reflect current military technologies, we might ask, "what is it like to be a drone operator?"–to which we are left with an answer that is truly unexpected.

To be a drone operator is to be prone to severe psychological harm. Such psychological harm might turn into physical harm. Thus, sitting in an air-conditioned cubicle thousands of kilometers away from the place where military operations take place does not guarantee safety. In fact, quite paradoxically, it threatens with the opposite. These "cubicle warriors" sometimes face difficulties nobody else faces. They sit in one place, but their minds extend to completely different places, where they have to witness and do things radically different from what they experience when they are away from work. Accordingly, their lives are full of dissonances.

There are also discussions about the presence or lack of martial courage in drone operators who control their war machines from afar (cf. Kirkpatrick 2015a, b; Sparrow 2015). It is undeniable that the extended mind theory perspective on drone operators, as proposed above, will shed new light on these discussions. It has the potential to redefine debates in this field, and also it might help to establish the (professional) identity of drone operators.

The area of the extended mind and its application to drone operators is a new research topic and so far open for more detailed theoretical and empirical studies. This chapter is a very first attempt to explain intricate aspects of this relatively new profession made possible by the technology of late twentieth and predominantly twenty-first century. My proposed explanation raises many further questions. For example, if drone operators suffer mentally, is it possible that their emotions are extended as well? It seems very likely. There are many other similarly interesting and important questions that invite us to conduct further research in this application of the extended mind theory.

Acknowledgement This article was produced with the financial support of the Ministry of Education, Youth and Sports within the National Sustainability Programme I, project of Transport R&D Centre (LO1610), on the research infrastructure acquired from the Operation Programme Research and Development for Innovations (CZ.1.05/2.1.00/03.0064).

References

Adams, F., & Aizawa, K. (2010). *The bounds of cognition*. Oxford: Wiley-Blackwell.

Armour, C., & Ross, J. (2017). The health and well-being of military drone operators and intelligence analysts: A systematic review. *Military Psychology, 29*(2), 83–98.

Asaro, P. (2013). The labor of surveillance and bureacratized killing: New subjectivities of military drone operators. *Social Semiotics, 23*(2), 196–224.

Carter, A., & Palermos, O. (2016). Is having your computer compromised a personal assault? The ethics of extended cognition. *Journal of the American Philosophical Association, 2*(4), 542–560.

Chamayou, G. (2015). *Drone theory*. London: Penguin.

Chappelle, W., Goodman, T., Reardon, L., & Thompson, W. (2014a). An analysis of post-traumatic stress disorder in United States Air Force drone operators. *Journal of Anxiety Disorders, 28*(5), 480–487.

Chappelle, W., McDonald, K., Prince, L., Goodman, T., Ray-Sannerud, B., & Thompson, W. (2014b). Symptoms of psychological distress and post-traumatic stress disorder in United States Air Force "drone" operators. *Military Medicine, 179*(8), 63–70.

Clark, A. (2007). Re-inventing ourselves: The plasticity of embodiment, sensing, and mind. *Journal of Medicine and Philosophy, 32*(3), 263–282.

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. Oxford: Oxford University Press.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7–19.

Crilly, T. (2010). UN official: Drone attacks controlled away from battlefield may lead to 'Playstation' mentality. *The Telegraph.* Accessed 15 Sept 2017. https://www.telegraph.co.uk/news/worldnews/northamerica/usa/7800586/UN-official-drone-attacks-controlled-away-from-battlefield-may-lead-to-PlayStation-mentality.html.

Cullen, T. (2011). *The MQ-9 Reaper remotely piloted aircraft: Humans and machines in action* (Doctoral dissertation). Retrieved from https://dspace.mit.edu/bitstream/handle/1721.1/80249/836824271-MIT.pdf.

Dao, J. (2013). Drone pilots are found to get stress disorders much as those in combat do. *The New York Times*. Accessed 19 Sept 2017. http://www.nytimes.com/2013/02/23/us/drone-pilots-found-to-get-stress-disorders-much-as-those-in-combat-do.html.

Dougherty, M. (2015). *Drones: An illustrated guide to the unmanned aircraft that are filling our skies*. London: Amber Books.

Draper, M. (2011). *Sitting ducks and peeping toms: Targets, drones and UAVs in British military service since 1917*. Leigh: Air-Britain.

Gal, S., Shelef, L., et al. (2016). The contribution of personal and seniority variables to the presence of stress symptoms among Israeli UAV operators. *Distaster and Military Medicine, 2*(18), 1–8.

Gusterson, H. (2015). *Drone: Remote control warfare*. Cambridge: MIT Press.

Hall, J. (1984). *American kamikaze*. Titusville: Del-Mar.

Hawkes, R. (2015) Post-traumatic stress disorder is higher in drone operators. *The Telegraph*. Accessed 19 Sept 2017. http://www.telegraph.co.uk/culture/hay-festival/11639746/Post-traumatic-stress-disorder-is-higher-in-drone-operators.html.

Kirchhoff, M. (2011). Extended cognition and fixed properties: Stepts to a third-wave version of extended cognition. *Phenomenology and the Cognitive Sciences, 11*(2), 287–308.

Kirkpatrick, J. (2015a). Drones and the martial virtue courage. *Journal of Military Ethics, 14*(3–4), 202–219.

Kirkpatrick, J. (2015b). Reply to Sparrow: Martial courage – Or merely courage? *Journal of Military Ethics, 14*(3–4), 228–231.

Kreuzer, M. (2016). *Drones and the future of air warfare: The evolution of remotely piloted aircraft*. New York: Routledge.

Lee, C. (2017). *Jet Lag*. New York: Bloomsbury Academic.

Levy, N. (2007). Rethinking neuroethics in the light of the extended mind thesis. *The American Journal of Bioethics, 7*(9), 3–11.
Martin, M., & Sasser, W. (2010). *Predator: The remote-control air war over Iraq and Afghanistan: A pilot's story*. Minneapolis: Zenith Press.
McCammon, S. (2017). The warfare may be remote but the trauma is real. *National Public Radio*. Accessed 19 Sept 2017. http://www.npr.org/2017/04/24/525413427/for-drone-pilots-warfare-may-be-remote-but-the-trauma-is-real.
McCurley, M., & Maurer, K. (2015). *Hunter-killer: Inside America's unmanned air war*. New York: Dutton.
Menary, R. (2008). Cognitive integration and the extended mind. In R. Menary (Ed.), *The extended mind*. Cambridge: MIT Press.
Mindell, D. (2015). *Our robots, ourselves: Robotics and the myths of autonomy*. Cambridge: MIT Press.
Nagel, T. (1974). What is it like to be a bat? *Philosophical Review, 83*(4), 435–450.
Power, M. (2013). *Confessions of a drone warrior*. GQ. Accessed 17 Sept 2017. https://www.gq.com/story/drone-uav-pilot-assassination.
Richardson, L., Frueh, C., & Acierno, R. (2010). Prevalence estimates of combat-related PTSD: A critical review. *The Australian and New Zealand Journal of Psychiatry, 44*(1), 4–19.
Rogers, A., & Hill, J. (2014). *Unmanned: Drone warfare and global security*. London: Pluto Press.
Royakkers, L., & Van Est, R. (2010). The cubicle warrior: The marionette of digitalized warfare. *Ethics and Information Technology, 12*(3), 289–296.
Rupert, R. (2009). *Cognitive systems and the extended mind*. Oxford: Oxford University Press.
Singer, P. (2010). *The soldiers call it war porn*. Der Spiegel. http://www.spiegel.de/international/world/interview-with-defense-expert-p-w-singer-the-soldiers-call-it-war-porn-a-682852.html.
Sparrow, R. (2009). Predators or plowshares? Arms control of robotic weapons. *IEEE Technology and Society Magazine, 28*(1), 25–29.
Sparrow, R. (2015). Martial and moral courage in teleoperated warfare: A commentary on Kirkpatrick. *Journal of Military Ethics, 14*(3–4), 220–227.
Vanhoenacker, M. (2015). *Skyfaring: A journey with a pilot*. London: Penguin.
Wagner, W. (1982). *Lightning bugs and other reconaissance drones*. Fallbrook: Aero Publishers.
Wallace, D., & Costello, J. (2017). Eye in the sky: Understanding the mental health of unmanned aerial vehicle operators. *Journal of Military and Veteran's Health, 25*(3), 36–41.
Whittle, R. (2014). *Predator: The secret origins of the drone revolution*. New York: Henry Holt.

Marek Vanžura PhD, is the head of the autonomous driving department at CDV – Transport Research Center, Czech Republic. He conducts research on autonomous vehicles in the form of self-driving cars and remotely operated vehicles. His specialization is in human-machine interaction, extended and distributed cognition, and the embodied mind. He holds a PhD in theory and history of science from Masaryk University, Brno.

Chapter 11
Extending Introspection

Lukas Schwengerer

The Case of Extended Introspection[1]

Let us start with a story about Otto:

> Otto suffers from Alzheimer's disease, and like many Alzheimer's patients, he relies on information in the environment to help structure his life. Otto carries a notebook around with him everywhere he goes. When he learns new information, he writes it down. When he needs some old information, he looks it up. For Otto, his notebook plays the role usually played by a biological memory. In short, Otto has beliefs extending to the notebook. One day I meet Otto in the streets of New York. Out of curiosity I ask him "Otto, I know where the museum is, but do you believe the museum is on 53rd street?" Otto looks at his notebook, finds the right entry, and answers "I believe the museum is on 53rd Street."[2]

At face value in this story Otto self-ascribes a belief state by looking at the notebook. But this leads straight into conflicting intuitions:

- On one hand Otto appears to simply detect his belief. The story is set up in a way that he has beliefs extending to the notebook, so the way to detect these beliefs is

[1] Special thanks to Giada Fratantonio, Grace Helton, Jesper Kallestrup, Aidan McGlynn and two anonymous reviewers for feedback on earlier versions of this paper. Large parts were presented at the University of Edinburgh WIP seminar and the conference 'Minds, Selves and 21st Century Technology' at the New University of Lisbon. Further thanks go to the audiences at these events.

[2] This is different from the extended self-knowledge case by Carter and Pritchard (2018), which has the self-ascription written down in the notebook: "[…] For example, when he learns new information about his own mental states (i.e., beliefs, feelings, desires, etc.) – information about his mental states which would be lost in biological storage – he writes it down in the notebook. Likewise, when he needs some old information about his mental life, he looks it up. For Otto*, his notebook plays the role usually played by a biological memory in preserving a mental narrative." I will use 'extended introspection' instead of 'extended self-knowledge' to avoid confusion.

L. Schwengerer (✉)
Philosophy Institute, University of Duisburg-Essen,
Essen, North Rhine-Westphalia, Germany
e-mail: Lukas.Schwengerer@uni-due.de

R. W. Clowes et al. (eds.), *The Mind-Technology Problem*, Studies in Brain and Mind 18, https://doi.org/10.1007/978-3-030-72644-7_11

to look at the notebook. And intuitively, what else is introspection other than directly detecting your own beliefs?
- On the other hand Otto appears to base his self-ascription on evidence that I, standing next to Otto, can use just as well to ascribe the belief to Otto. And this is certainly not what we intuitively think introspection is like.

With these conflicting intuitions at hand, we need to search for a proper way to understand extended introspection. Of course, one way out would be to simply deny the extended mind thesis. However, I will be working with the assumption that the extended mind thesis is true and explore the implications of the thesis for self-knowledge. Given this assumption Otto definitely self-ascribes a mental state by looking at the notebook. But what kind of self-ascription are we dealing with here? This is the core question I am going to address in this paper. I start by stepping back and taking a quick overview on the notion of extended belief as the basis for extended introspection. I then consider the standard options for self-ascribing mental states: introspection and self-directed mind-reading. I discuss these in turn and show that both are not a great fit. This leads into a dialectic dilemma with no clear way out. I then propose a unique account of extended introspection based on epistemic rules, inspired by Alex Byrne's (2005) account of introspection. On my proposal extended introspection turns out to be reliable, because self-verifying under extended belief condition. Moreover, I show in the final section that my solution for the case of Otto and his notebook is also applicable to twenty-first century versions of the case, relying on smartphones, wearable technology and the internet.

Extended Belief

My story about Otto introspecting extended beliefs builds on the Otto case presented by Clark and Chalmers (1998).[3] The central idea behind ascribing Otto plus notebook an extended belief lies in the fact that the notebook plays the role usually played by biological memory. This idea features in the argument as the parity principle:

> *Parity Principle*: If, as we confront some task, a part of the world functions as a process which, were it to go on in the head, we would have no hesitation in recognizing as part of the cognitive process, then that part of the world is (so we claim) part of the cognitive process (Clark and Chalmers 1998, p. 8).

The notebook does the job of the biological memory, because Otto consistently uses it just like we usually use our biological memory. Whenever he gets new information he writes it down. Whenever he wants to remember something he looks it up in the notebook. And the notebook is with him all the time. So why not straight up accept the notebook as Otto's memory?

The difference in location is not a well-motivated rationale to dismiss the similarities, if we follow Clark and Chalmers. However, this needs a further, more

[3] Clark (2010) explicitly allows that the 'reading in' of information from devices might count as introspection depending on how one classifies the overall case. He states that "[f] rom our perspective the systemic act is more like an act of introspection than one of perception" (Clark 2010, p. 56).

detailed look. Even if we follow the parity principle, we need to say something more on the notebook's role in Otto's cognitive processes. Not every notebook makes a good case for something that "we would have no hesitation in recognizing as part of the cognitive process," that is in this case, as part of memory. I personally use notebooks every once in a while. But if you were to observe my sporadic looks at the notes, you would surely hesitate to call my notebook part of my memory. Moreover, sometimes I write notes so illegibly that there is no way to decipher the content later on. Other times I find myself disagreeing with my notes – "I can't have meant that," I say to myself. And finally, often my notebook is just not with me, whereas my biological memory is consistently with me. Or at least I cannot remember the cases in which it was not. In short: My notebook's role is very differently from biological memory. We have to be careful not to trivialize extended beliefs in a way that obscures their differences with non-extended beliefs.

Clark and Chalmers are fully aware of this concern. Therefore they propose four conditions that external aids have to satisfy to play a role in an extended belief (Clark and Chalmers 1998; Clark 2010).

External aids must be:

(i) consistently available,
(ii) readily accessible,
(iii) automatically endorsed, and
(iv) present because they were consciously endorsed in the past.

They are slightly skeptical about the previous endorsement condition, claiming that this requirement is debatable. The appeal of (iv) is that by getting rid of it one might lose grip of the difference between remembering and relearning. However, insofar as (iv) requires conscious endorsement, which is an internal condition, (iv) seems to rob the extended belief thesis of its core motivation. The cognitive work would not be extended enough, so to speak (cf. Bernecker (2014, p. 5)). For my purposes condition (iv) is not important, so I leave the debate open. However, conditions (ii) and (iii) will become important later on. Fortunately, they are generally accepted as requirements for extended beliefs.

Clark (2008) and Menary (2007) introduce additional conditions of two-way interactions. The idea is that the interactions between the external aid and the person have to be connected in such a way that they continuously affect the other. Clark calls this *continuous reciprocal causation*, Menary *cognitive integration*. An old-fashioned, paper-based book, for instance, would not fit this criterion because the reader does not affect the book sufficiently. For my purposes the conditions (i) to (iii) above are enough, so I can stay neutral on any further requirements. I will also remain neutral on whether the analogy to memory works as well as Clark and Chalmers want it to. The use of notebooks can fail more frequently and in different ways than our biological memory, which might undercut the parallel drawn by the argument for the extended mind (cf. Rupert (2004), Sterelny (2004)). However, for the present purpose of discussing extended introspection, I will assume that the case can be made in favor of Clark's and Chalmers's picture. Hence, my approach can be seen as a discussion under a conditional: What should we think of introspection of

extended beliefs, *if there are extended beliefs and they can be roughly characterized by conditions (i) to (iii).*

There is a final worry to address concerning the focus on the simple Otto case. The case is highly idealized. Why should we care about this case at all? Perhaps looking at other, more realistic scenarios requires different conditions or dimensions of integration (as discussed in Sutton (2006), Sutton et al. (2010), Sterelny (2010), Menary (2010), and Heersmink (2015))? After all, today's technological gadgets are more complex than the simple notebook. My first response here is to point out that the case does not strike me as unrealistic, even though it involves an uncommon situation. Moreover, there are sufficiently similar cases featuring individuals and smartphones or smartwatches that satisfy the same conditions (i) to (iii). For these cases my discussion should still be applicable. I am going to discuss one paradigmatic case more in depth towards the end, but I am now going to address the idea that similar cases are plausible featuring twenty-first century technology. Fortunately, such a scenario has already been provided by Paul Smart (2018). He presents the following case Otto++:

> Otto++ is a neurologically impaired individual who is biologically identical to Otto. Otto++ has just purchased a shiny new smartphone and augmented reality glasses. Otto++ spends some time configuring his phone by installing a variety of apps. He then carries his phone with him wherever he goes. In order to ensure that he has access to relevant information, Otto++ installs an app that enables him to record important pieces of information. The app, however, does not store information locally on the device. Instead, it relies on a semantically enabled, cloud-based personal data store that stores information in a linked data format. In order to access his personal data store, Otto++ installs an app that enables him to quickly retrieve important items of information using an intuitive graphical user interface. He also links his phone to his augmented reality glasses so that relevant information from his data store can be presented within his visual field. One day, while on a trip to New York City, Otto++ decides he would like to visit MoMA. He automatically says the word "MOMA" out loud. His phone executes a semantic query against his personal information repository and retrieves information about MoMA. A set of directional indicators appear within Otto++'s visual field, alongside some descriptive information about MoMA (Smart 2018, p. 279).

Smart argues that Otto++ still satisfies conditions equivalent to my (i) to (iii). To show that we have to unpack the story a little. One important difference to the notebook case is that Otto++ does not store information on any particular device. Otto++ stores the information on a cloud-based personal data store that can be accessed with various different devices. He carries some of these devices with him, but he could in principle also use a friend's smartphone if needed. Moreover, the information is stored in a specific format[4] such that it can be retrieved quickly and easily based on queries. This is meant to counteract the difficulties of accessing a specific piece of information in a large database. These features combined allow Otto++ to have the information constantly available and readily accessible, at least to a degree similar to biological memory. Hence, the case passes (i) and (ii). Condition (iii) appears to be satisfied as well. There is nothing in the case that prevents Otto++

[4] For details on the format see Smart (2018, pp. 273-274).

from automatically endorsing the information retrieved from his personal cloud-based data store. We can even stipulate that (iv) is satisfied. Simply suppose that information can only be uploaded to the data store by Otto++ himself and (iv) is met. Given these similarities Smart suggests that "[…] the case of Otto++ is sufficiently similar to the original case […]" (Smart 2018, p. 279). I agree. Hence, we can imagine an Otto++ case for extended introspection as well. Otto++ apparently can come to know what he believes by suitably attending to his cloud-based data store. Moreover, I seem to be able to come to know what Otto++ believes by suitably attending to Otto++'s cloud-based data store.

My plan is to focus on the simpler case of Otto and his notebook and then discuss whether introspection in the Otto++ case can be explained by the same strategy as the simpler case. The focus on the notebook case makes the discussion a little easier, without significant drawbacks. Nevertheless, it is important to point to potential problems in generalizing from the simple case to more complex cases, such as Otto++. I will respond to these problems towards the end. However, I will not discuss speculative cases involving belief constituting cognitive enhancements that are directly connected to the brain. I find these scenarios still rather alien and it is difficult to have clear intuitions about them.

11.1 Extended Introspection as Introspection

To find out whether the extended introspection case is a genuine case of introspection, we need to define our criterion for such genuine introspection. A good starting point for doing so is to look at our intuitive judgments about introspection, and most importantly what they are contrasted to. The two important points of contrast are first, knowledge of the non-mental world; and second, knowledge of other people's mental states. Introspection appears to be different to both. Self-knowledge seems to some extent *privileged* and *peculiar* (Byrne 2005). It is privileged insofar as one's beliefs about one's own mental states are more likely to amount to knowledge than one's beliefs about other's mental states or the external world. Moreover, they are peculiar, insofar as they are formed by a special method or way of knowing. Call this the *cognitive access view* of self-knowledge. Cognitive access accounts come in different shapes. For instance, they can accept a peculiar detectivist method of introspection (e.g. Armstrong (1968), Nichols and Stich (2003), Goldman (2006), Macdonald (2014))[5] or an empiricist transparency story as in Byrne (2005), and Fernández (2013).

The cognitive access view is not the only description of self-knowledge available. For instance, s*elf/other parity* accounts of self-knowledge argue that the specialness of self-knowledge is overstated. Self-beliefs are largely[6] formed by the

[5] These lists are not exhaustive.

[6] 'Largely' because they usually limit the parity to propositional attitudes.

same processes we use to attribute mental states to others (cf. Carruthers (2011), Cassam (2014)). Moreover, a*gentialist* (sometimes called *rationalist* (Gertler 2011)) positions (e.g. Burge (1996), Moran (2001), Bilgrami (2006)) understand privilege and peculiarity not in terms of better access, but with a particular first-personal connection between agents and their mental states. For Moran (2001) this particular connection is also the basis for the transparency of beliefs, that is, that one can know whether one believes that *p* by attending to the question of whether *p* is true. Agentialist accounts accept that self-belief is special because it is about *my* mental states, and I am responsible for *my* beliefs. Self-knowledge in this conception is important as a precondition for critically reflecting on one's own mental states.

For this paper I will restrict myself to the cognitive access view due to limited space. Hence, I understand privileged access for introspection as a person being more reliable in self-ascribing mental states than in ascribing mental states to other people. Moreover, I take peculiar access to denote some sort of peculiar way of knowing one's own mental states, compared to knowing other people's mental states ('mind-reading').[7]

With the features of privileged access and peculiar access as the defining features for introspection we can start looking at whether extended introspection presents us with these features. When Otto self-ascribes his belief by looking at the notebook, does he have privileged and peculiar access?

The first step to investigate his privileged position is to consider whether Otto is reliable. Clearly he is in the way the story is set up. Otto has, by stipulation, the extended belief that the museum is located on 53rd street. When I ask him, he looks at the notebook and avows that he believes the museum is on 53rd street. And in doing so, he makes a correct statement. It is true that he has this belief. Even more so, it seems very difficult for Otto to be wrong about himself in this case. Given that the notebook is consistently available, readily accessible, and automatically endorsed, Otto is highly reliable in looking at the notebook and self-ascribing the belief. In nearly all nearby possible worlds in which Otto looks at the notebook to self-ascribe the belief, he will read the notebook correctly, understand what is written in the notebook and successfully attribute the belief to himself. If this was not the case, then Otto would fail to have the appropriate connection to the notebook and not have an extended belief. But I stipulated the extended belief conditions to be satisfied. Hence Otto has to be reliable.

However, is Otto *more* reliable than other people looking at his notebook? Privileged access requires more than just reliability. It requires being more reliable than other people in attributing mental states to Otto. Consider the following take on Otto:

[7] Moreover, I also bracket discussions of the relation of Moore's paradox to self-knowledge, even though they are often given prominent space. For a sample of views on Moore's paradox see Green and Williams (2007). For recent discussions of Moore's paradox and self-knowledge see Shoemaker (1995), Kind (2003), Williams (2004), Fernández (2013), Coliva (2016) and Smithies (2016).

> Otto suffers from Alzheimer's disease, and like many Alzheimer's patients, he relies on information in the environment to help structure his life. Otto carries a notebook around with him everywhere he goes. When he learns new information, he writes it down. When he needs some old information, he looks it up. For Otto, his notebook plays the role usually played by a biological memory. I look at Otto's notebook and read that the museum is at the 53rd street. Therefore, I judge that Otto believes the museum is at 53rd street.

Is my belief that Otto believes the museum is at 53rd street reliably formed? This depends on how we individuate the process. If the process in question is "looking at notebooks" in general, then it probably is not. However, given that I stipulated Otto's extended belief, it seems the answer is a clear yes if we individuate the process based on looking at *Otto's* notebook. That is, if I ascribe Otto beliefs based on looking at *his* notebook, I will be highly reliable. As long as I can read Otto's handwriting and understand his language, I will end up with true beliefs about Otto's extended beliefs.[8] Therefore I seem to be equally reliable as Otto in ascribing extended beliefs to Otto. Otto is not privileged.

Perhaps I am a little too quick here. One might bring up at least two objections. First, one may claim that there is a sense in which Otto is privileged. Otto has his notebook with him all the time, whereas I don't have the same kind of permanent access. So he has better access after all! To this I respond that this is not the right sort of privilege that I am talking about. It is important to distinguish two kinds of privileged access here. First, one can have 'accidentally' privileged access because one is more often in possession of the relevant evidence, but that evidence (or sufficiently similar evidence) is in principle accessible for other people. Second, one can have 'essentially' privileged access, when the evidence (or sufficiently similar evidence) is not in principle accessible for other people. Otto seems to have accidentally privileged access to his extended belief, but no essentially privileged access. And it is the latter that is of interest here. If I were to follow Otto all the time and look at his notebook constantly there would not be an extended belief of his that he knows but I do not.

Second, one can argue that one needs to have reasons to base judgements of Otto's beliefs on the notebook. One needs to know not only that this is Otto's notebook, but also that Otto is related to the notebook in a way that satisfies the conditions for extended belief. I can accept the former, but deny the latter. One needs some reason to ascribe a belief based on the notebook, but knowing that this is Otto's notebook is enough.[9] After all, Otto himself does not need to know that the extended belief conditions are met. He just needs to look at the notebook and self-ascribe the belief. The same applies to me ascribing mental states to Otto by looking at his notebook. I need to have some reason why this notebook relates to Otto's

[8] And there seems little reason to give Otto any privilege with reading and understanding the notebook, or at least nothing that could not be removed by modifying the case slightly.

[9] This includes knowing something about the function of a notebook. Some knowledge of the function is necessary to distinguish one's notebook from a scrap of paper lying around. That it is my notebook and not a random scrap of paper provides a prima facie reason to attribute to myself the content. Thanks to Grace Helton for pointing out this issue.

beliefs, but those reasons need not entail any of Otto's beliefs. These reasons are only necessary to prompt me to ascribe a belief on the basis of the notebook. They need not be reasons that guarantee the truth of my ascription.

What about peculiar access? Is there anything special about Otto's self-ascription? At first glance what Otto does and what I do when we ascribe a belief based on the notebook seems not that different. We both look at the very same thing and use it as a basis for a belief ascription. It is the same notebook. Sure, he has access to the notebook constantly, but that does not change the fact that when we both look at it to ascribe a mental state, we do the same thing. Otto might use the notebook for much more, but for this single belief-forming process it is hard to find a difference. That the notebook constitutes his first-order belief does not influence the second-order belief formation. Moreover, there is no reason to assume more inferential work being done by me than Otto. Sure, I need to recognize that this is Otto's notebook. But so does Otto. I don't see any reason why I would need any additional inferential step that Otto does not need. I can treat whatever is written in the notebook as direct evidence for Otto's belief, just as Otto does. Gertler (2007) uses this observation to argue against the existence of extended beliefs. As I will show later this is not the route that I want to take. However, Gertler is right that this symmetry between Otto and me is bad news for peculiar access and overall bad news for extended introspection as a subspecies of genuine introspection.

11.2 Extended Introspection as Mind-Reading

Mental state ascriptions are commonly assumed to be either introspectively formed, or based on mind-reading.[10] I just argued that extended introspection does not fit the criteria of privileged and peculiar access, so it does not look like it goes into the introspection category. One should then expect a better fit with mind-reading. And if so, calling it 'extended introspection' might actually be misleading. To see whether this is true I want to start with a rough and ready characterization of mind-reading. I understand mind-reading here as a capacity to attribute mental states to human beings based on behavioral observations and evidence of the situation/environment. To illustrate this take this simple mind-reading story by Jordi Fernández (2013, p. 4):

> Suppose that one of the things that you believe about me is that I want Barcelona FC to win the UEFA Champions League. Suppose, furthermore, that your belief is justified. What could justify your belief? Perhaps you heard me express that desire, or you observed me screaming at the TV while we watched one of the Champions League games, or you noticed my mood when I read in the news that the team was not doing so well in that competition. (Fernández 2013, p. 4)

[10] Some argue for a distinct method of knowing other people's mental states that is not inferential mind-reading (cf. Spaulding (2015)). I will not consider these options.

Whatever can justify your belief about his mental state has to be something you observed. You cannot directly access his mental state, but rather you need to base it on the evidence you gather by perception. You can listen to his testimony and you can see him get emotional when watching the game. Perhaps even facial expressions showing his mood can be sufficient if you know enough about the situation he is in right now. Crucially, the mental state attribution is based on things other than the mental state attributed. Plausibly they are not completely unrelated – his emotional reaction is connected to the desire that Barcelona FC wins – but they are different things. This is not to say that mental states cannot play any role in mind-reading processes. Rather, the mental state that should be attributed at the end of a process cannot itself be the input of the very same process. If I ascribe mental state M_1 by process P_1 at t_1 I can later on use M_1 in process P_2 at t_2 to ascribe M_2. For instance, I ascribe the desire for Barcelona FC to win the Champions League to Jordi Fernández based on his behavior while watching a football game. Then later on I see him celebrating after the game finished. I can now use the previously ascribed desire to figure out that he is happy that Barcelona FC won.

Mind-reading approaches that start with behavioral observation plus evidence of the situation come in two varieties: As theory-theory accounts (e.g. Gopnik and Wellman (1994, 2012), Gopnik and Meltzoff (1997)) and as simulation accounts (e.g. Goldman (2006)). Both differ in what is done with the observational input, but for my purpose the focus is on the input itself. So I can work with a very simplified black-box model.

The model can be self-directed, so it is possible to self-ascribe a mental state based on behavioral observation and a grasp of the current situation. A standard example for this type of case is Wright's explanation of a scene in Jane Austen's Emma:

> Emma has just been told of the love of her protégée, Harriet, for her — Emma's — bachelor brother-in-law, a decade older than Emma, a frequent guest of her father's, and hitherto a stable, somewhat avuncular part of the background to her life. She has entertained no thought of him as a possible husband. But now she realizes that she strongly desires that he marry no one but her, and she arrives at this discovery by way of surprise at the strength and color of her reaction to Harriet's declaration, and by way of a few minutes' reflection on that reaction. She is, precisely, not moved to the realization immediately; it dawns on her as something she first suspects and *then* recognizes as true. It *explains* her reaction to Harriet. (Wright 1998, pp. 16–17)

With self-directed mind-reading in the mix, how can we tell a mind-reading story of Otto's self-ascription? Otto simply looks at the notebook and avows that he believes the museum is at 53rd street. There is no strong, colorful reaction to the notebook that Otto then can interpret in such a way that makes it possible to attribute a belief. The only reaction is the assertion that he believes that the museum is at the 53rd street. But that reaction already presupposes what the mind-reading process wants to get at. It is useless as a basis for a mind-reading process. What other behavior can be considered as the basis for Otto? Perhaps he can base his mind-reading on previous instances of looking at the notebook. In the past he read the notebook and acted accordingly. But is this enough as a basis? I doubt it. While he might get to the general conclusion that usually he acts according to what is written in the notebook, there is no way this general claim can lead him to the specific attribution of a belief *that p*. Where should the propositional content come from, if his basis is a general claim?

The obvious amendment is to let the content written in the notebook play a role in the explanation. So the general observation of Otto's past behavior plus the fact that p is written in the notebook are the input for the mind-reading process. The output is then the self-ascription that he believes that p – that the museum is at 53rd street in this case.

But this is no good either. There are red flags for both parts of the input. First, it does not seem obvious whether Otto can use his past behavior at all as an input. Otto has Alzheimer's after all. How can we let any representation of his past behavior play a crucial role in the belief-production if it is unclear whether his memory supports any such representation? If Otto needs a notebook to remember where the museum is, he likely won't be able to remember how he behaved in the past. We can imagine him writing down all his behavior in the past as well, but that seems unnecessary in our story about Otto's self-ascription. The story is complete without him also checking the notebook for his previous behavior.

Second, if we accept that p written in the notebook plays a role as input, we might be already stepping away from mind-reading. Remember that the mental state attributed at the end of the mind-reading process cannot be already the input. If we let the notebook play such a pivotal role as input, we are in danger to abandoning this general rule, because the notebook stating that p is part of the extended belief that Otto wants to self-ascribe. This means that Otto would effectively use his extended belief to self-ascribe the very same belief. While this perfectly fits the story, it does not tell a mind-reading tale anymore. Instead we have to deal with Otto directly detecting his own extended belief. He bases his second-order belief on his first-order belief. Suddenly the mind-reading approach no longer seems appropriate.

The proponent of the mind-reading solution can attempt to save his account by implementing a different move. It is not the present behavior that Otto interprets, but rather his past behavior. That there is something written in the notebook is evidence of his own past action of writing down the location of the museum. When Otto recognizes what is written in the notebook, he recognizes that this is something that he wrote and hence, that he believes. In this case he bases his self-ascription on the evidence of his prior action. The first problem with this solution is that it heavily

relies on Otto having written that *p* into the notebook earlier. However, as I noted before, the previous endorsement condition is not all that necessary according to Clark and Chalmers. And if there is no previous endorsement condition, then there can be cases in which the writing in the notebook is no evidence of Otto's past actions. Second, in some cases Otto will not be able to tell whether he wrote *p* into the notebook. Just think of notes on a smartphone. There is no handwriting that can be recognized as something Otto himself wrote. Nevertheless, he can have beliefs extended to the notes on the smartphone and be unable to rule out that someone else wrote them. It seems that he does not need to know who put the notes in there. If this is correct, then he cannot use these notes as results from his past actions. Hence, he cannot use them as a basis for self-directed mind-reading.

I showed that both the introspection and the mind-reading approach fail to capture extended introspection appropriately. What are we supposed to do now? I believe there are four options available, with little to go in favor of each.

1. We can change our account of introspection, such that privileged- and peculiar access are less important.
2. We can change our account of mind-reading, such that we can allow the ascribed belief to already be a part of the input in some sense.
3. We can get rid of the idea of extended belief altogether, and stick to our guns for introspection and mind-reading.
4. We can propose that extended introspection needs its own, distinct account of producing self-knowledge.

It is a difficult dialectic position to be in. Moreover, if one looks at the individual debates, then one can find independent motivation for every single of these options. One can accept that introspection does not come with this sort of privileged- and peculiar access in general with Carruthers (2011), Cassam (2014), or Schwitzgebel (2008) and go for (1). One can adapt ideas from direct social perception of mental states such as presented in Krueger (2012) or Spaulding (2015) and go for (2). One can get rid of extended minds with Gertler (2007) or Adams and Aizawa (2010) and go for option (3). I, however, want to opt for (4). The reasoning is largely defensive. If one chooses option (4), one need not change any independently motivated position. Therefore, (4) is the least invasive way to go. I can avoid the internal debates of introspection, mind-reading, and extended mind for the bargain of accepting a new source of self-knowledge: Extended Introspection.

11.3 Extended Introspection Sui Generis

Perhaps it should not surprise us that both introspection and mind-reading do not work for our story about Otto. Why expect a phenomenon to fit an explanation that was modeled after very different cases? So let's start looking at the case and build an account from the bottom up.

The main idea is to describe what is going on in the Otto case and then transform this description into a general principle or rule. This approach is not entirely original. Rather, it is a staple in the epistemologist's toolbox to build epistemic rules out of cases, where an epistemic rule is simply a rule of belief formation.

For instance, we can start with the following story taken from Alex Byrne (2005 p. 93):

> Mrs. Hudson might hear the doorbell ring, and conclude that there is someone at the door. By hearing that the doorbell is ringing, Mrs. Hudson knows that the doorbell is ringing; by reasoning, she knows that there is someone at the door.

This case is straightforward. Mrs. Hudson believes that someone is at the door, because the doorbell rings. We can transform this into a rule that Mrs. Hudson follows:

DOORBELL If the doorbell rings, believe that there is someone at the door.

It is easy to see that this rule fits the case. Mrs. Hudson's belief formation can be described as her following this conditional, whereas following the conditional means that she forms the consequent belief *because* she recognizes that the antecedent condition holds. Generalizing this, Byrne (2005, p. 94) states that S follows the Rule R ('If conditions C obtain, believe that *p*') on a particular occasion iff on that occasion:

(a) S believes that *p* because she recognizes that conditions C obtain

Which implies:

(b) S recognizes (hence knows) that conditions C obtain
(c) Conditions C obtain
(d) S believes that *p*

DOORBELL happens to be a good rule, that is, it usually produces true beliefs. On the other hand, you can think of bad rules that produce false beliefs most of the time. For instance, "If you are hungry, believe that it is sunny outside" is a rule that will not generate true beliefs in general. There are simply too many instances in which you are hungry, but it is not sunny outside. In other words, the rule is unreliable.

I can now make use of epistemic rules to get a grasp of what goes on in the Otto case. The case was the following:

> Otto suffers from Alzheimer's disease, and like many Alzheimer's patients, he relies on information in the environment to help structure his life. Otto carries a notebook around with him everywhere he goes. When he learns new information, he writes it down. When he needs some old information, he looks it up. For Otto, his notebook plays the role usually played by a biological memory. I ask Otto whether he believes that the museum is on 53rd Street. Otto looks at his notebook and answers "I believe the museum is on 53rd Street."

Just as I described Mrs. Hudson's belief formation, I can now describe Otto's belief formation as a two-step process. Otto looks at his notebook, and then self-ascribes a belief because of what is written in the notebook. He recognizes that the notebook

says that *p*, and transitions, by reasoning, to the belief that he believes that *p*. He thereby fits the following rule:

NOTEBOOK If your notebook says that *p*, believe that you believe that *p*.

Otto believes that he believes that *p*, because he recognizes that his notebook says that *p*. This looks quite similar to Byrne's (2005) general rule BEL: If *p*, believe that you believe that *p* (Byrne 2005, p. 95). There is a sense in which it is just an instance of BEL, because Otto recognizes that *p* by looking at the notebook. However, I do not opt to use BEL here. My rationale is twofold. First, BEL shows a very strong asymmetry between a first person formulation and a third person version. We already saw that the case of extended introspection does not fit with this asymmetry to such an extent. NOTEBOOK on the other hand captures the close similarity to a third-person rule that fits our initial intuition that extended introspection is not quite as privileged and special as genuine introspection. Second, NOTEBOOK can provide additional insights to the Otto case. Both reasons will become apparent in this section.

So far NOTEBOOK looks just like DOORBELL. However, it is more than the simple DOORBELL rule. NOTEBOOK is a very special rule, if one supposes that the conditions for extended beliefs are satisfied for Otto. Remember conditions (ii) and (iii) that external aids have to satisfy for an extended belief. The external aid has to be (ii) readily accessible, and (iii) automatically endorsed. (ii) plays an important role in making it possible to follow NOTEBOOK. Epistemic rules in general do not say whether they can actually be followed. DOORBELL, for instance, gives you a conditional that provides a transition from recognizing the doorbell ringing to a belief that will likely be true, if the antecedent holds. But you might not be able to recognize that the doorbell is ringing. You could be deaf, or simply listening to music on headphones on a volume that makes it impossible to hear the doorbell. Nothing guarantees that a good rule is one that you can follow. However, this is different for extended believing Otto and the NOTEBOOK rule. Otto is guaranteed to be able to follow NOTEBOOK reliably with regard to recognizing the antecedent.[11] The argument for this is rather simple:

1. Otto has a belief extending to the content of his notebook. (Assumption)
2. Otto's belief can only be extended if the content of the notebook is readily accessible. (Conditions for Extended Belief)
3. Otto can readily access the content of the notebook (from 1, 2)
4. If the content of the notebook is readily accessible, then Otto can reliably recognize that the notebook says that *p*, if it says that *p*. (Spelling out Accessibility)

[11] It is important to highlight the difference between a rule being reliable and one being able to follow the rule reliably. The rule is reliable (good) if it mostly produces true beliefs. On the other hand, one can follow an inferential rule reliably if one usually is in a position to follow it. One can reliably follow a reliable rule, but one can also reliably follow an unreliable rule. The same goes for being unable to reliably follow a rule.

5. Otto can reliably recognize that the notebook says that *p*, if it says that *p*. (from 3, 4)

The argument shows that Otto is able to reliably recognize that the antecedent of the NOTEBOOK rule holds, if it holds. He can do so in virtue of the extended belief condition that makes the notebook readily accessible. The only step in the argument that is not an assumption or already independently argued for is (4), but I take this to be intuitively true. What else could it mean to be able to readily access a notebook, if not that I can reliably recognize that the notebook says that *p*, if it does say that *p*?[12]

That Otto can reliably recognize that the notebook says that *p* is not enough to guarantee that he can reliably follow NOTEBOOK completely. He further needs to be able to follow the conditional and form a belief according to the conditional. This is not worrisome at all. As long as Otto is able to reason, he is able to follow a conditional just fine.

So far I established that Otto can reliably follow NOTEBOOK, given that he has beliefs extending to his notebook. However, I still need to provide reasons why NOTEBOOK is actually a good epistemic rule. Why should NOTEBOOK generate true rather than false beliefs? Here I take another condition of extended belief to play the pivotal role. This time it is condition (iii), the automatic endorsement condition. The idea is that whenever Otto looks into his notebook and reads that *p*, he automatically endorses that *p* and thereby is guaranteed to believe that *p*. This can be used in an argument as follows:

1. Otto automatically endorses that *p*, if he reads that *p* in his notebook. (Condition of Extended Belief)
2. Endorsing that p entails believing that *p*. (Spelling out Endorsement)
3. Otto reads that *p* in his notebook. (Assumption)
4. Otto endorses that *p*. (from 1, 3)
5. Otto believes that *p*. (from 2, 4)
6. If Otto reads that *p* in his notebook, he believes that *p*. (from 3, 4, 5)

This conclusion shows that NOTEBOOK is actually a good epistemic rule. It is good, because the consequent belief will always be true when Otto follows the rule. Whenever Otto follows NOTEBOOK, he starts by looking at the notebook which says that *p*. Otto recognizes that *p* and automatically endorses it. This endorsement guarantees that he believes that *p*. So when Otto follows the conditional and forms the belief that he believes that *p*, he will be correct. Following NOTEBOOK is infallible, because the mere act of following the rule guarantees the second-order belief to be true by securing the first-order belief.

[12] In earlier versions I was tempted to read the ready access condition stronger than merely reliable access. However, that would be against Clark and Chalmers (1998) intention of providing a parallel to biological memory. Clark (2010) uses reliable access instead of ready access to avoid this confusion. Thanks to Brie Gertler for pointing this out.

A crucial step in the argument is (2), which depends on the notion of endorsement in play. Clark and Chalmers (1998) do not provide much information in this regard. However, Clark (2010) says that endorsement means that "It should not usually be subject to critical scrutiny (unlike the opinions of other people, for example). It should be deemed about as trustworthy as something retrieved clearly from biological memory" (Clark 2010, p. 46). I take this to entail belief, insofar as it is equal to regarding *p* as true. If Otto recognizes that the notebook says that *p*, he holds *p* to be true without additional, critical scrutiny. He will use *p* as a premise in practical and theoretical reasoning, the same as if he would hold *p* to be true based on any other source. This should be uncontroversial, given that I assume that the extended mind thesis is true.

Behind this argument lies a general observation of the Otto case. If Otto self-ascribes a belief by looking at the notebook two usually[13] unrelated factors coincide. On one hand there is the ground for a belief. For instance, I can form a perceptual belief based on a perceptual seeming. I form the belief that the sun is shining, because I have a visual experience of the sun shining. This is my evidence that I base my belief on. However, on the other hand there is an external fact that makes the belief true. My belief that the sun is shining is true, if the sun is in fact shining. This truthmaker is different from the basis of my belief. In the Otto case things seem different. The very same thing that Otto bases his second-order belief on also makes this second-order belief true. He looks at the notebook and believes that he believes that *p* because of the notebook saying that *p*. At the same time the notebook makes it the case that he believes that *p*, thereby making his second-order belief true. This makes it difficult for Otto to be wrong about himself, if he self-ascribes by looking at the notebook. The external aid is both the basis and the truthmaker for his self-ascription. Here even the most determined sceptic cannot find a gap in which to insert his knife – as long as the sceptic is on board with extended beliefs in general.

I established that NOTEBOOK is an epistemic rule that Otto can reliably follow, and moreover a good, truth-conducive rule. However, where does NOTEBOOK put extended introspection with respect to privileged and peculiar access? To answer this I want to look at third person equivalents to the NOTEBOOK rule. After all, the prior intuition was that there is nothing special about Otto looking at the notebook compared to me looking at the notebook. So perhaps there is a similar epistemic rule for me. And I believe there is, let's call it O-NOTEBOOK.

O-NOTEBOOK If Otto's notebook says that *p*, believe that Otto believes that *p*.

Under the assumption that Otto has beliefs extended to the notebook, O-NOTEBOOK also looks like a very good rule. If I look at Otto's notebook and recognize that it says that *p*, then it will be true that Otto believes that *p*. It will be true, because by the assumption of extended beliefs whatever is written in the notebook constitutes dispositional beliefs of Otto, just the same way something stored

[13] Even though not always. One might argue that the same thing happens if I form beliefs about my qualia.

in biological memory would. However, there are some differences to NOTEBOOK. First, O-NOTEBOOK plus the assumption that Otto has extended beliefs does not guarantee that one can reliably follow the rule. Whereas Otto can reliably follow NOTEBOOK in virtue of the extended mind conditions, there is no condition that guarantees me any access to Otto's notebook. Hence, I might not be able to recognize the antecedent of the conditional in a large number of cases.

Second, Otto can ascribe *occurrent* beliefs by using NOTEBOOK, whereas O-NOTEBOOK cannot do the same. The idea here is that whenever Otto looks at his notebook to follow his NOTEBOOK rule he thereby endorses the content right at that moment. That is, the endorsement, and thereby the belief that *p*, plays an active role in Otto's cognitive machinery at the moment of him following NOTEBOOK. Moreover, it is plausible that Otto will be consciously aware of his belief that *p*, when he follows NOTEBOOK. On the other hand, if I look at Otto's notebook, there is no way for me to tell whether Otto believes that *p* occurrently or dispositionally. It is possible that I ascribe to Otto the belief that *p* by looking at his notebook, while at that moment Otto himself does not look at the notebook at all and is thinking about something completely unrelated to *p*. In this case, I can still correctly ascribe a belief that *p* to Otto, but only a dispositional belief. The difference can be spelled out by expanding both epistemic rules:

NOTEBOOK* If your notebook says that p, believe that you occurrently believe that p.

O-NOTEBOOK* If Otto's notebook says that p, believe that Otto occurrently believes that p.

NOTEBOOK* is a good rule, whereas O-NOTEBOOK* is not. The former will produce mostly (always) true beliefs, but the latter generates a ton of false beliefs in cases where I follow O-NOTEBOOK* when Otto does not look at his notebook. Hence there is a difference between NOTEBOOK and O-NOTEBOOK insofar as they both use a general notion of 'belief', but have different types of beliefs as truth-makers in general.

With these differences in mind I can confidently say that there is something peculiar and special about extended introspection. But it is only a minor difference. Nothing guarantees that I can reliably follow the third person equivalent to Otto's NOTEBOOK rule. Moreover, even when I can follow O-NOTEBOOK, I cannot employ quite the same method as Otto. However, I can do something in the vicinity, closely resembling Otto's belief formation. And I can be reliable as well; I am just limited in the range of reliable extended mental state ascriptions. I cannot reliably ascribe occurrent states based on Otto's external aids, but I can reliably attribute his extended beliefs in general. The result is somewhere in between the features that ordinary self-knowledge and mind-reading are said to possess. It is not quite as peculiar as self-knowledge based on usual introspection, but there is still some difference between the first-personal access and the third-personal one.

From Otto to Otto++

So far I have discussed the extended introspection case for Otto and his notebook. Earlier on I remarked that the same strategy that works as an explanation for extended introspection in the case of Otto and his notebook also works for Smart's (2018) case of Otto++, who has information stored online on a personal cloud service that can be accessed via smartphones, smartwatches or augmented reality glasses. Otto++ has sufficient access to the information via these devices to satisfy conditions (i) to (iii). Otto++ poses two interesting challenges. First, can we come up with an epistemic rule for extended introspection that fits the case? And second, is there still the same kind of privilege involved in this case?

To answer the first question, I propose to use an analog to the NOTEBOOK rule:

CLOUD If your cloud-based data store says that *p*, believe that you believe that *p*.

We have to be liberal in interpreting the phrase "if your cloud-based data store says […]." The idea is that this phrase covers all cloud-based information being presented by any device that is sufficiently connected to that cloud-based store. The antecedent can be satisfied if your smartphone app connected to the cloud-based data store shows you that *p*, but also if the augmented reality glasses tell you that *p*, or a device produces sounds providing you with the information that *p*. In all of these cases the same arguments I provided for the notebook case can be applied in relation to the CLOUD rule. If you have beliefs extended to the cloud-based data store satisfying conditions (i) to (iii), you will be in a position to reliably follow the rule CLOUD. Moreover, you will automatically endorse the content provided by the cloud-based data store and therefore believe that *p* whenever you follow the rule.

Other people will also be able to come to know Otto++'s beliefs by accessing his cloud-based data store. So we get the third-person rule for Otto++:

O-CLOUDIf Otto++'s cloud-based data store says that *p*, believe that Otto++ believes that *p*.

However, just as with O-NOTEBOOK, there is no guarantee that I am able to reliably follow O-CLOUD. On the other hand, the conditions for extended belief guarantee that Otto++ can reliably follow CLOUD. Hence, Otto++ is privileged with regard to his cloud-based beliefs just like Otto is with regard to his notebook-based beliefs.

In addition, Otto++ is also privileged with regard to his occurrent beliefs. To see this, consider the rule for occurrent beliefs:

CLOUD* If your cloud-based data store says that *p*, believe that you occurrently believe that *p*.

For the same reasons as in NOTEBOOK* the rule CLOUD* will be a good rule and produce true beliefs, whereas the third-person equivalent O-CLOUD* will not.

O-CLOUD* If Otto++'s cloud-based data store says that *p*, believe that Otto++ occurrently believes that *p*.

Just as in the notebook case, CLOUD* is a good rule, but O-CLOUD* is not. Following O-CLOUD* does not reliably lead to true beliefs. Hence, Otto++ is also privileged with regard to the occurrent nature of his extended beliefs. Overall, Otto++ has privileged access to his extended beliefs similar to Otto in the notebook case, hence I found my answer the first question.

To answer the second question, I want to consider the following scenario. Suppose that the FBI can observe Otto++ constantly by utilizing the technological gadgets Otto++ uses. The augmented reality glasses have accessible cameras built in that are always filming Otto++ and the FBI can use his smartphone to listen in on what he says at all times. Finally, suppose that they use well developed computer programs and well trained agents to infer Otto++'s mental states from the observed behavior. In this scenario the FBI seems to be in an especially good position to know Otto++'s mental states. With all these surveillance tools the FBI might be able to get to know Otto++ even better than Otto++ knows himself. Is this a problem for the claim that extended introspection is privileged?[14] I do not think so. I am willing to grant that someone could come to know *more* about Otto++'s beliefs than he himself knows. And the constant observation via wearable technology seems to be a plausible scenario in which that is the case. However, this does not threaten the privileged access Otto++ has to his extended and non-extended beliefs. We should not understand privileged access as having an especially high number of justified true beliefs about one's mental states. Rather, following Byrne (2005), I suggest that we should understand privileged access as beliefs formed by introspection being especially likely to amount to knowledge. There is less of a chance for making a mistake when introspecting. This is also the case for extended introspection. CLOUD is a rule that is self-verifying under extended belief condition. If Otto++ follows CLOUD he will end up with a true belief. Moreover, the conditions for having beliefs extended to the cloud-based data store guarantee that Otto++ can reliably follow the rule. Nothing guarantees that anyone else can follow the third-person rule O-CLOUD. One might respond here that O-CLOUD does not seem to be a difficult rule to follow, especially for someone who is able to observe everything that Otto++ does at any given time. But even in that case, at best one could be on par with Otto++ with regard to his dispositional beliefs. Otto++ is still privileged with regard to occurrent beliefs formed by CLOUD*. Otto++'s observable behavior will merely provide you with a reasonable basis to infer Otto++'s occurrent beliefs, but this basis underdetermines his actual mental states. Any behavior that might indicate that Otto++ likely accesses information (e.g. all the mechanical steps that lead to the information being displayed plus his eye movement) could happen without any information being accessed. Otto++ might look as if he accesses information from the cloud when he actually does not. Hence, there is no third-personal path to know whether Otto++ occurrently believes that p that can compete with Otto++'s privileged access. Otto++ can use CLOUD* and thereby accurately find out that he occurrently believes that p. Hence, there is still room for Otto++'s privileged access.

[14] Thank you to a reviewer who raised this issue.

11.4 Conclusion

I showed that a straightforward interpretation of extended introspection as genuine introspection is not appropriate. Furthermore, a self-directed mind-reading story does not fit either. This left us with four distinct options. We can change our accounts of introspection and mind-reading, deny the extended-mind thesis, or propose a sui generis form of extended introspection. I chose the latter. Therefore, I proposed an original account of extended introspection based on epistemic rules. I provided a rule for the Otto example, and argued that this happens to be a rule that Otto can follow and that generates true beliefs. Both are secured by the requirements of extended beliefs. If Otto has extended beliefs, then he can reliably follow the NOTEBOOK rule which leads him to true beliefs about his mental states. Finally, I showed how this picture fits the ideas of privileged and peculiar access. Otto is privileged, because having beliefs extended to the notebook (plus reasoning) guarantees that he can reliably follow the NOTEBOOK rule. Furthermore, he is in a special position because he can attribute that a belief is occurrent based on NOTEBOOK, whereas a third-person variation of the rule, O-NOTEBOOK, cannot do the same. It is still an open question to what extent this result can generalize to other cognitively integrated artifacts. This is mainly a question of whether (i)-(iii) hold for all extended mental states, or whether some artifact can be integrated enough to count as part of the mind without satisfying (i)-(iii). Plausibly, widespread technology such as smartphones and smartwatches fit these conditions to some extent. In particular I argued that it does generalize to Smart's (2018) Otto++ case, which satisfies (i) to (iii). The other open question is how extended beliefs can be combined with an agential account of self-knowledge. I leave this to future work.

References

Adams, F. R., & Aizawa, K. (2010). Defending the bounds of cognition. In R. Menary (Ed.), *The extended mind* (pp. 67–80). Cambridge, MA: MIT Press.

Armstrong, D. M. (1968). *A materialist theory of mind*. London: Routledge/Kegan Paul.

Bernecker, S. (2014). How to understand the extended mind. *Philosophical Issues, 24*, 1–23.

Bilgrami, A. (2006). *Self-knowledge and resentment*. Cambridge, MA: Harvard University Press.

Burge, T. (1996). Our entitlement to self-knowledge I. *Proceedings of the Aristotelian Society, New Series, 96*, 91–116.

Byrne, A. (2005). Introspection. *Philosophical Topics, 33*(1), 79–104.

Carruthers, P. (2011). *The opacity of mind: An integrative theory of self-knowledge*. Oxford: Oxford University Press.

Carter, A. J., & Pritchard, D. (2018). Extended self-knowledge. In J. Kirsch & P. Pedrini (Eds.), *Third-person self-knowledge, self-interpretation, and narrative* (pp. 31-49). Berlin: Springer.

Cassam, Q. (2014). *Self-knowledge for humans*. Oxford: Oxford University Press.

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. Oxford: Oxford University Press.

Clark, A. (2010). Memento's revenge: The extended mind, extended. In R. Menary (Ed.), *The extended mind* (pp. 43–66). Cambridge, MA: MIT Press.

Clark, A., & Chalmers, D. J. (1998). The extended mind. *Analysis, 58*(1), 7–19.
Coliva, A. (2016). *The varieties of self-knowledge*. London: Palgrave Macmillan.
Fernández, J. (2013). *Transparent minds: A study of self-knowledge*. Oxford: Oxford University Press.
Gertler, B. (2007). Overextending the mind? In B. Gertler & L. Shapiro (Eds.), *Arguing about the mind* (pp. 192–206). New York: Routledge.
Gertler, B. (2011). *Self-knowledge*. New York: Routledge.
Goldman, A. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mind-reading*. Oxford: Oxford University Press.
Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts and theories*. Cambridge, MA: MIT Press.
Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. Hirschfield & S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 257–293). New York: Cambridge University Press.
Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanism and the theory theory. *Psychological Bulletin, 138*(6), 1085–1108.
Green, M. S., & Williams, J. N. (Eds.). (2007). *Moore's paradox: New essays on belief, rationality, and the first-person*. Oxford: Oxford University Press.
Heersmink, R. (2015). Dimensions of integration in embedded and extended cognitive systems. *Phenomenology and the Cognitive Sciences, 13*(3), 577–598.
Kind, A. (2003). Shoemaker, self-blindness and Moore's paradox. *The Philosophical Quarterly, 53*(210), 39–48.
Krueger, J. (2012). Seeing mind in action. *Phenomenology and the Cognitive Sciences, 11*(2), 149–173.
Macdonald, C. (2014). In my 'Mind's Eye': Introspectionism, detectivism, and the basis of authoritative self-knowledge. *Synthese, 191*(15), 3685–3710.
Menary, R. (2007). *Cognitive integration: Mind and Cogntition unbounded*. Basingstoke: Palgrave Macmillan.
Menary, R. (2010). Dimensions of mind. *Phenomenology and the Cognitive Sciences, 9*(4), 561–578.
Moran, R. (2001). *Authority and estrangement*. Princeton: Princeton University Press.
Nichols, S., & Stich, S. (2003). *Mindreading: An integrated account of Pretence, self-awareness, and understanding other minds*. Oxford: Oxford University Press.
Rupert, R. (2004). Challenges to the hypothesis of extended cognition. *Journal of Philosophy, 101 (8)*, 389–428.
Schwitzgebel, E. (2008). The unreliability of naive introspection. *Philosophical Review, 117*(2), 245–273.
Shoemaker, S. (1995). Moore's paradox and self-knowledge. *Philosophical Studies, 77*, 211–228.
Smart, P. (2018). Emerging digital technologies: Implications for extended conceptions of cognition and knowledge. In A. J. Carter, A. Clark, J. Kallestrup, O. S. Palermos, & D. Pritchard (Eds.), *Extended epistemology* (pp. 266–304). Oxford: Oxford University Press.
Smithies, D. (2016). Belief and self-knowledge: Lessons from Moore's paradox. *Philosophical Issues, 26*, 393–421.
Spaulding, S. (2015). On direct social perception. *Consciousness and Cognition, 36*, 472–482.
Sterelny, K. (2004). Externalism, epistemic artefacts and the extended mind. In R. Schantz (Ed.), *The externalist challenge* (pp. 239–254). Berlin: De Gruyter.
Sterelny, K. (2010). Minds: Extended or scaffolded? *Phenomenology and the Cognitive Sciences, 9*(4), 465–481.
Sutton, J. (2006). Distributed cognitions: Dimains and dimensions. *Pragmatics and Cognition, 14*(2), 235–247.
Sutton, J., Harris, C. B., Keil, P. G., & Barnier, A. J. (2010). The psychology of memory, extended cognition, and socially distributed remembering. *Phenomenology and the Cognitive Sciences, 9*(4), 521–560.

Williams, J. N. (2004). Moore's paradoxes, Evans's principle and self-knowledge. *Analysis, 64*(4), 348–353.
Wright, C. (1998). Self-knowledge: The Wittgensteinian legacy. In C. Wright, B. Smith, & C. Macdonald (Eds.), *Knowing our minds* (pp. 13–45). Oxford: Oxford University Press.

Lukas Schwengerer is a postdoctoral researcher at the University of Duisburg-Essen. He received his PhD from the University of Edinburgh in 2018 with a thesis on a unified transparency account of self-knowledge. He works primarily on questions in epistemology and philosophy of mind.

Chapter 12
Epistemic Complementarity: Steps to a Second Wave Extended Epistemology

Gloria Andrada

12.1 Introduction

Our lives are permeated by technology. We are currently living at a time of unprecedented technological mediation: a radical transformation of how we think, act and relate to each other. Take the production of this chapter as an example. It is the result of a long chain of bodily movements and sensorimotor encounters with technological devices and informational resources: reading essays, taking notes in notebooks, writing with laptops, late-night searches of the World Wide Web, etc. The list could go on and on.

According to the thesis of *extended cognition,* some of the artifacts we pervasively interact with are, under certain circumstances, literally part of our cognitive processes.[1] *Extended cognition* promotes an egalitarian approach to cognition and invites us to see the material realizers of cognition as encompassing not just activities of the brain and body, but also the activities of some elements of the organism's external environment. In doing so, it challenges the long-assumed *intracranial* view, and submits that cognition trespasses the old bounds of skin and skull. For the purposes of this chapter, I will assume that extended cognition is not only a genuine possibility but is also instantiated in many actual cases.[2]

During the last 10 years, the epistemological implications of extended cognition have received increasing attention, giving rise to the now flourishing project of *extended epistemology*.[3] The recognition of our ability to couple with external

[1] Clark and Chalmers 1998, Clark 2008.

[2] For discussion, see Menary 2010.

[3] See especially the essays in Carter et al. 2018b.

G. Andrada (✉)
Instituto de Filosofia da Nova, Faculdade de Ciências Sociais e Humanas Universidade Nova de Lisboa, Lisbon, Portugal
e-mail: gandrada@fcsh.unl.pt

R. W. Clowes et al. (eds.), *The Mind-Technology Problem*, Studies in Brain and Mind 18, https://doi.org/10.1007/978-3-030-72644-7_12

devices to accomplish a variety of cognitive tasks has reshaped the debate concerning our epistemic dependence on artifacts, that is, our dependence on artifacts to achieve epistemic goals (e.g. knowledge, justified beliefs) or to extend our cognitive abilities (Pritchard 2010).

Despite the progress that has been made in this field, I believe that standard extended epistemology has led to an inadequate framework for investigating the type of practices required for making extended cognitive processes epistemically good. The principal aim of this chapter is to show why this is the case, and to make room for an alternative, more attractive framework.

In particular, most writings in the field of extended epistemology have built their extended epistemology from a conception of extended cognition modeled from a *first-wave approach*.[4] This approach is characterized by the quest for functional parity between intracranial cognitive processes and extended cognitive ones. This, in turn, has led to what I will call an *epistemic parity* approach to epistemic hygiene. By this I mean, roughly, that whatever makes intracranial cognitive processes epistemically benign, also works for extended cognitive processes. This leads to a stringent account that forestalls any opportunity of capturing the complexity of the epistemic standing of our pervasive interactions with technologies, and the type of individual engagement required for being healthy epistemic agents.

The model I will propose for extended epistemology is built from a *second-wave extended cognition* approach. According to this route to cognitive extension, extended cognitive processes are characterized by the complementary functionalities that they bring to purely intracranial ones. It is not a matter of parity, but rather complementarity and the integration of quite heterodox elements that coordinate towards accomplishing cognitive tasks.

Building on the importance of this complementary relation between purely organic faculties on the one hand, and cultural and technological artifacts and practices on the other, I will take the first steps towards a new framework for extended epistemology. The main idea is that determining what is required for achieving epistemic hygiene in extended cognitive processes will ultimately depend on the complex interplay between the diverse embodiments of knowers and the salient properties of technological artifacts, as well as the socio-cultural environment in which the interaction is embedded. By attending to these three aspects, we will begin the complex task of analyzing the impact of new technologies on our individual cognitive and epistemic capabilities. This work aims to provide a unifying framework, guided by what I will call an epistemic complementarity principle.

The plan for the paper is as follows. First, I briefly introduce two lines or agendas behind extended cognition (Sect. 12.2). Then I present the central tenet behind extended epistemology, based on an epistemic parity principle (Sect. 12.3). In contrast with this, I introduce an epistemic complementarity principle and provide a new model for extended epistemology (Sect. 12.4). I finish the chapter by applying this model to a case where an agent interacts with a smartwatch (Sect. 12.5).

[4] Clark and Chalmers 1998, Clark 2008.

12.2 Waves of Extended Cognition

According to extended cognition theory, elements external to the organism can, under certain circumstances, participate in the mechanistic realization of cognitive states and processes (Clark and Chalmers 1998; Clark 2008). Following John Sutton's (2010) categorization, there are different lines or agendas behind extended cognition.

First-wave extended cognition relies on the now-famous *Parity Principle* (Clark and Chalmers 1998). The Parity Principle is a heuristic for detecting cognitive extension, motivated by a common-sense functionalist approach to cognition.[5] The heuristic is simple: if an external process (one that includes elements that trespass the organic boundaries) is such that, were it to happen inside the head, we would consider it *cognitive*, drawing from our common-sense or folk knowledge about cognition, then that process *is* itself cognitive. The original Parity Principle was formulated as follows:

> If, as we confront some task, a part of the world functions as a process which, were it to go on in the head, we would have no hesitation in accepting as part of the cognitive process, then that part of the world is part of the cognitive process. (Clark and Chalmers 1998, p. 8)

The parity principle played a historically important role in rejecting the unprincipled exclusion of external objects as parts of cognitive processes. Once such discrimination was removed, it became possible to look for criteria of cognitive extension. To this end, Clark and Chalmers (1998) themselves advocated an approach which emphasizes an object's portability, general availability and constancy across contexts. The resulting criterion is captured by the so-called "glue and trust" conditions (Clark 2010), namely: availability, more-or-less automatic endorsement, and easy access.

Indeed, it appears that what is crucial for cognitive-extending technologies is that they be available, and that they can easily enter an agent's cognitive routines as part of their problem-solving machinery. These criteria capture different functional profiles of cognitive states and processes, drawing from our common-sense knowledge about intracranial cognition. The idea is that candidates for genuine cognitive extension should be available when needed, easily or directly accessed and automatically endorsed, to approximately the same extent as intracranial cognitive abilities.

Andy Clark and David J. Chalmers's proposals have sparked critical discussion from a number of theorists, which have eventually resulted in the development of so-called s*econd-wave extended cognition* (Sutton 2010, pp. 193–201).

The main tenets of the s*econd-wave* approach are driven by a need to move beyond using coarse-grained functional similarities to characterize cognitive extension.[6] The main resistance to the parity line of thinking emerges from the

[5] For defences of this argument, see Clark and Chalmers 1998, and Clark 2008, and for recent rejections of this argument see Sprevak 2009, and Wadham 2015.

[6] Menary 2006, 2007, 2010; Rowlands 2010, Sutton 2010, Sutton et al. 2010, Heersmink 2014, Kiverstein and Farina 2011.

homogenization of inner and outer capabilities, in virtue of the fact that inner and outer resources are heterogeneous in functionally relevant ways.

Building on the work of Merlin Donald (1991), John Sutton articulates the rationale behind extending cognitive processes (memory in particular) in terms of the complementary contributions afforded by their different functionalities.

The complementarity thesis for cognitive extension is captured in the following C*omplementarity Principle*:

> In extended cognitive systems, external states and processes need not mimic or replicate the formats, dynamics or functions of inner states and processes. Rather, different components of the overall (enduring or temporary) system can play quite different roles and have different properties while coupling in collective and complementary contributions to flexible thinking and acting. (Sutton 2010, p. 194)

This principle emphasizes how diverse functionalities across different resources in fact explain the purpose of cognitive extension, since they complement each other. For instance, the typical features of *exograms* or external symbols are quite different from *engrams* or the brain's memory traces (Donald 1991, p. 308; Sutton 2010, p. 189). The former usually last longer, are more easily transmissible across media and contexts, and can be retrieved and manipulated by a greater variety of means (Donald 1991, pp. 315–316).[7]

The challenge is to investigate how these diverse elements integrate with each other to achieve a given cognitive task. In this framework, extension is viewed as a continuous and fuzzy phenomenon, rather than an all-or-nothing matter. The idea is that the level of integration among heterodox resources will vary depending on the agent's cognitive profile and the technology's properties. Higher degrees of integration will entail that they are genuine parts of a cognitive system, while lesser degrees will be seen as a symptom that they are not parts but rather tools or scaffolds.[8]

Once the emphasis on coarse-grained functional similarities is abandoned, room is made for investigating the embodied engagements in virtue of which cognition is extended (Rowlands 2009; Menary 2007). Cognition is extended through the sensorimotor manipulation of external resources, including external representations.[9] Importantly, as Richard Menary has argued, such embodied manipulations are embedded in a wider social, semantic, and normative environment (Menary 2010, p. 11). These manipulations are governed by social practices, some of which are cognitive in nature (Menary 2007, 2018a). It is through processes of enculturation that we get to be readers, writers, smartphone users and web surfers. This means that in order to explain and understand extended cognition, we need to look at the

[7] The properties of exograms are not fixed and might in fact change, depending on their format and implementation.

[8] See Heersmink 2014 for a thorough taxonomy of the dimensions of integration, and Sterelny 2010 for an account of scaffolded cognition.

[9] Rowlands 2009, Menary 2007, 2010, 2018a, b.

complementary interplay between organisms, the technological resources they interact with, and the socio-cultural environment in which they are embedded.[10]

The preceding remarks were intended to illustrate the main tenets of the first- and second-wave approaches to extended cognition. Now, I will show their impact on extended epistemology. Crucially, while second-wave extended cognition provides a well-established framework for modeling extended cognition, its bearing on extended epistemology remains to be explored. In the following sections, I aim to show that it provides the building blocks for a different and more promising way of thinking about the epistemic standing of contemporary and emerging technologies.

12.3 Epistemic Parity

Extended epistemology is concerned with studying the consequences that the program of extended cognition has on our epistemic evaluations. Traditionally, epistemology has taken an intracranial perspective while considering what makes a process a cognitive process. The limits of epistemic agents have thus been determined by their organic boundaries. That is why the program of extended cognition and its extension of the individual agent has led to a fascinating discussion in epistemology, giving rise to the project of extended epistemology.

An analysis of the literature reveals that most accounts of extended epistemology have, almost unreflectively, been built on a first-wave approach to extended cognition.[11] Accordingly, they model cognitive (and epistemic) extension in terms of coarse-grained functional similarities. Moreover, most approaches to extended epistemology have reinforced the importance of similarities between intracranial and extended cognitive processes. This has led theorists to endorse or advocate an epistemic parity principle such as the one proposed by J. Adam Carter (2013), who formulates it as follows:

> E-Parity Principle: For agent *S* and belief *p*, if *S* comes to believe *p* by a process which, were it to go on in the head, we would have no hesitation in ascribing knowledge of *p* to *S*, then *S* knows *p*. (Carter 2013, p. 4203)

The idea is that extended cognition should avoid not only any bio-prejudice concerning what does or does not count as the physical machinery of a cognitive process, but also any epistemic bio-prejudice. According to an epistemic bio-prejudice,

[10] Elaborating on the social and cultural dimensions of cognition has led to what can be identified as *third-wave extended cognition* (sketched in Sutton 2010, and developed in Kirchhoff 2012, Kirchhoff and Kiverstein 2019, Gallagher 2013, and Cash 2013). The active role of the socio-cultural environment is also captured in the epistemic complementarity approach I present in this paper, giving rise, perhaps, to a third-wave extended epistemology. I plan to return to this in future work.

[11] See Carter, Palermos, Kallestrup and Pritchard 2014, and Carter 2017. Notice that the conditions for evaluation will depend on our epistemological commitments, hence giving rise to different kinds of epistemic parity principles. See e.g. Carter 2017, pp. 9–12.

the difference between intracranial and extracranial cognitive processes can be interpreted as an epistemic difference (Carter 2013, pp. 4202–4203). For simplicity's sake, the principle is framed in terms of knowledge. However, epistemic parity affects other epistemic domains as well, insofar as cases of extended cognition are subject to ascriptions of "justification, understanding, rationality or intellectual virtue" (ter 2013, p. 4202).

While not every account within the emerging project of extended epistemology has explicitly endorsed an epistemic version of the parity principle, most of them have taken for granted that extended cognitive processes and intracranial ones must be similar from the standpoint of our epistemic evaluations. To see this, it will be useful to introduce the notion of epistemic hygiene (Clark 2015; Carter et al. 2014).

To get an initial handle on the notion of epistemic hygiene, we might draw an analogy with sanitary hygiene. Sanitary hygiene is a set of practices and standards aimed at preserving health and preventing the spread of diseases (cf. Nicolle 2007). Similarly, the notion of epistemic hygiene captures the idea that there are some practices required for our individual and collective epistemic well-being. An agent needs to acquire good habits in order to be epistemically healthy. These might include, for instance, checking or taking care of the reliability of one's belief-forming methods. An *epistemic disease* might be understood as the spreading of unreliable methods for producing beliefs, or the acquisition of bad habits that lead to more false beliefs than true ones. Notice that the analogy also holds concerning the harm of excessive (epistemic) hygiene. Excessively hygienic practices might turn out to be detrimental to one's epistemic well-being, just as excessive hygiene is detrimental to the general health of an organism.[12]

Since there is such a thing as epistemic hygiene, we need to establish what kind of engagement is required for an agent to be epistemically hygienic. Given that offering a complete account of these practices goes beyond the purpose of this chapter, we must focus on a minimal requirement. Epistemic hygiene is closely related to the reliability of our knowledge-producing methods and to the preservation of this reliability. We can thus take the reliability (truth-trackingness) of one's belief-forming processes, and some form of reaction to the shifting reliability of one's belief-forming processes, as minimal requirements of epistemic hygiene.[13]

Drawing from the work developed in Clark (2015), we can distinguish two ways of pursuing this minimal epistemic hygiene:

1. *Active pursuit of epistemic hygiene*: Clark (2015) refers to an "active pursuit epistemic hygiene" in relation to those practices involving the agents *themselves*, that is, practices that involve "person-level" engagement. Practices of this sort include, for instance, recognizing the source of the reliability of the technological process, or consciously inspecting the method used to obtain a certain piece

[12] Let me clarify that I do not want to suggest that an agent's epistemic life can be reduced to their epistemic hygiene; rather this notion serves the purpose of illustrating the different ways of thinking about extended epistemology.

[13] For more on this see Palermos 2014.

of information. In the context of new technologies, this might entail consciously inspecting the reliability of a piece of equipment. This is a form of "deliberate, conscious, slow, careful, agentive attention" (Carter et al. 2018a, p. 333).

2. *Passive pursuit of epistemic hygiene*: On the other hand, passive epistemic hygiene concerns those practices that involve little or no person-level engagement by the epistemic agent. For instance, it may be the result of the correct functioning of biologically-endowed sub-personal mechanisms, or of the proceduralization of certain practices in the form of (unconsciously enacted) patterns of behavior (see Menary 2012).

It is important to remark that Clark (2015) does not explicitly refer to a passive pursuit of epistemic hygiene, however he does contrast an agentive form of epistemic hygiene with sub-personal forms of epistemic hygiene, where sub-personal mechanisms react to the shifting degree of reliability of different sources of information, without any type of agential engagement. To this extent, although the active/passive distinction is a bit schematic, it intuitively captures two different aspects of the pursuit of epistemic hygiene, and it allows us to illustrate the difference between an epistemic parity approach and the framework I will present.[14]

The question that concerns me here is what type of engagement is required for achieving a minimal form of epistemic hygiene in extended cognitive processes. I will show that an epistemic parity approach incorrectly limits the answer that can be given to this question.

According to an extended epistemology based on first-wave extended cognition, or motivated by the E-parity principle, extended cognitive processes are evaluated in parallel with unextended cognitive processes. Remember that this way of thinking about extended cognition focuses on the similarities between extended cognitive processes and intracranial ones. The idea is that the technology should be *as* easily accessed as the faculties that lie beneath the agent's skin. In other words, to borrow an expression from Pritchard (2018, p. 96), extended and intracranial processes should be "phenomenologically on par."[15]

This, in turn, has led to a sort of parity principle concerning epistemic hygiene, namely the view that in order to ensure that a cognitive process (be it extended or not) is minimally hygienic, the type of individual engagement required must be the same. This means that if intracranial processes are epistemically hygienic in a passive or sub-personal way, the same goes for extended cognitive processes.

For instance, Clark (2015) argues that a sub-personal form of epistemic hygiene is the only form of epistemic hygiene compatible with extended cognition. Putting more stringent constraints on extended cognitive processes, such as inspecting the technology, is seen as disrupting parity and thus going against the central tenets of

[14] The debate concerning epistemic hygiene is orthogonal to the more traditional debate concerning internalism vs. externalism about epistemic justification. However, it is true that active epistemic hygiene involves increasingly strict conditions, and these in turn might point to more internalistic aspects of epistemic justification.

[15] Cf. Smart 2018a.

extended cognition. After all, intracranial cognitive processes are not usually subjected to an active type of inspection.[16] Moreover, it can be seen as a manifestation of the feared bio-prejudice. In order to prevent this unwanted situation, Clark concludes that a sub-personal form of epistemic hygiene is the only form of epistemic hygiene compatible with extended cognition and thus extended knowledge. He does so by relying on the predictive processing framework and the powerful sub-personal mechanisms of precision-estimation.[17]

Despite the pioneering role and great interest of the epistemic parity approach, I believe that it prevents an adequate extended epistemology. No doubt the parity principle has played a seminal role in raising awareness of a diffuse bio-prejudice concerning the role of the external environment in cognition. Nevertheless, tracing the source of extension to coarse-grained similarities prevents us from duly taking into account the deep differences that lie between organic and technological elements, and as we will see, these differences matter epistemically. I will show that from an epistemic standpoint, taking the complementarity framework seriously entails that even if external resources are genuine parts of a cognitive system, the relevant types of epistemic hygienic practices might vary precisely because of their differences.

12.4 Epistemic Complementarity

In this section I will present the central tenets of an epistemic complementarity framework.

12.4.1 The Epistemic Complementarity Principle

An epistemic complementarity framework starts from the idea that, in order to fully understand what does it take to achieve epistemic hygiene in extended cognitive processes, we need to offer a careful analysis of the elements that constitute such processes. Drawing from the complementarity principle formulated by Sutton (2010), I will now introduce an epistemic complementarity principle which captures the main motivation for the framework for extended epistemology that I am presenting here.

> **Epistemic complementarity principle**: In extended knowledge cases, the agent needs not mimic or replicate the engagement required for achieving knowledge in virtue of inner states and processes. Rather, different hygienic epistemic practices might be required for flexible extended knowing.

[16] This reasoning is dramatized by the Epistemic Hygiene Dilemma in Carter et al. 2018a: 334, and Clark 2015: 3763. See also Andrada (2019) for a new solution to this dilemma.

[17] See Clark 2015, pp. 3768–3771.

The central idea behind this principle is that, given that artifacts and other external resources have different properties and functions that complement inner cognitive states and functions, the cognitive processes that involve them might in fact require different types of engagements to achieve epistemic hygiene. Contrary to the epistemic parity principle, it accepts that the differences between internal and external elements might also have an impact on their epistemic standing. For simplicity's sake, this principle is also framed in terms of extended knowledge.

At this point it might be helpful to remember that the central idea behind the complementarity route to cognitive extension is that embodied agents deploy the functional and informational properties of cognitive artifacts to complement their onboard cognitive capacities. Instead of focusing on coarse-grained similarities, emphasis is placed on individual differences (both the cognitive and embodied profile of the organism, and the properties of the artifact) in order to investigate the integration between them. Extended cognitive systems are investigated across dimensions of integration between the embodied agent and the technological artifact, and the wider socio-cultural environment in which their interaction takes place.[18]

The principle of epistemic complementarity goes one step further and states that, even when an artifact is genuinely integrated into an agent's cognitive system, we should not assume that minimal epistemic hygiene requires them to be engaged with it to the same extent that they are with their intracranial cognitive processes. For this reason, instead of seeking epistemic parity, the epistemic complementarity principle compels us to pause and look at the contributions of the different elements that make up an extended cognitive process. This is captured in the following three steps.

12.4.2 Three Steps

For the purpose of this chapter, I have assumed that a minimal form of epistemic hygiene in cognitive processes involves, on the one hand, the reliability (truth-trackingness) of one's belief-forming processes and, on the other hand, some form of reaction to the shifting degree of reliability of one's belief-forming processes. This can be pursued either *actively,* involving some form of awareness or agentive engagement, or *passively*, in a way that is more or less automatically given, as with the correct functioning of our endowed capacities.

According to the epistemic complementarity approach, in order to determine the sort of agentive engagement required for achieving a minimal form of epistemic hygiene, and thus for being a candidate for extended knowledge, we must look at the interplay between the different cognitive and embodied profiles of the cognitive agent and the salient properties of the technological artifacts, as well as the environment in which they are embedded. These elements will determine whether (and to

[18] Sutton 2006, 2010, Menary 2010, Heersmink 2014.

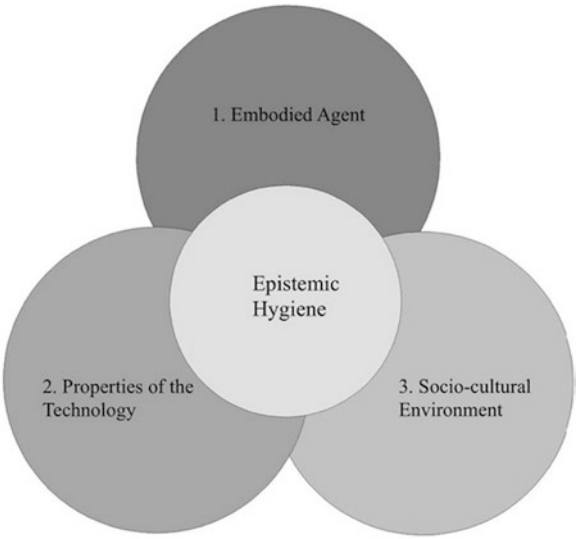


Fig. 12.1 Epistemic complementarity

what extent) epistemic hygiene in these contexts requires an active or passive pursuit, and the type of agentive engagement and effort required.

The moral is that we should not take for granted that extended cognitive processes require only whatever works for intracranial processes in order to be epistemically good. What the epistemic complementarity framework captures is the fact that epistemic hygiene is distributed, given that all the elements that constitute an extended cognitive process (embodied agent, technological artifact and sociocultural environment) have a bearing on the pursuit of epistemic hygiene. This can be represented in a threefold model (Fig. 12.1).

12.5 Step 1: Embodied Agents

The first step of an epistemic complementarity-based approach is to acknowledge the embodied dimension of extended cognition and extended knowledge. Our body is the interface that allows us to interact with different technologies. Whether this interaction takes place by swiping, touching a screen, or rotating lenses in order to see something, our body is a crucial element in our epistemic engagement with the world. As Robert Clowes has recently stated, we interact with contemporary technologies via *skilled gestures* (Clowes 2018, p. 4). Given this embodied dimension, an epistemic complementarity approach begins by looking at the diverse embodiments of knowers and their embodied and cognitive profiles. The question that needs to be answered is how an agent's embodiment might affect our account of epistemic hygiene.

The importance of diverse embodiments in epistemology has been emphasized by many feminist epistemologists, but has received little attention in extended epistemology, despite the crucial dimension of embodiment within extended cognition. This is why a distinction drawn by Antony (2002) is especially illuminating here.

According to Louise Antony (2002), a knower's embodiment might have two types of effects on epistemology: (i) *ground-level* effects and (ii) *meta-level* effects. The former effects capture the idea that diverse embodiments matter to how agents know. This might be settled partly by empirical studies, similar to those on cultural variation (Nisbett et al. 2001), or to those in the growing field of neurodiversity (Fenton and Krahn 2007). In contrast, meta-level effects focus on how a theorizer's embodiment affects how they theorize about epistemic matters. Such effects reflect how theorizing might depend on contingent and non-universal features of its practitioners' embodiments. Both levels are complex and deserve a deeper treatment than the one I will give here, but I will proceed to show how they relate to each other and their relevance for an extended epistemology.

An epistemic complementarity framework focuses mainly on ground-level effects: how an agent's cognitive profile and embodiment matter for how they acquire epistemic hygiene in extended cognitive processes. The motivation of such a focus derives from the complementarity route to extended epistemology, in combination with a meta-level reflection concerning how theorizers about extended cognition and extended epistemology (including myself) might have been driven by homogenizing assumptions and biases. Let me explain this further.

First of all, as I have just sketched, an epistemic complementarity approach acknowledges that there are relevant differences between extended and unextended processes, and this calls for taking into account the embodiment of potential extended knowers, as well as the bearing and effects that their different embodiments might have on the pursuit of epistemic hygiene. Second of all, a meta-level reflection motivates a study of the ground-level effects of the embodiment of extended knowers, because this helps us identify the different biases that might have guided our extended epistemology. For example, one might assume that all knowers are alike, and in doing so, one might take one's own particular way of experiencing one's cognitive and epistemic life (for instance, in relation to how one easily exercises one's own cognitive capacities) to be the standard or even the *only* way of doing so. Not only will this kind of strategy limit our own account of extended knowledge but it will also preclude us from grasping what it takes to achieve extended knowledge in real-life settings, as we will proceed to see.[19]

Concerning ground-level individual differences, little has been studied in relation to extended cognitive processes. This is surprising, since we find an appeal to individual differences, in particular with regard to personality psychology, in some of the canonical texts of second-wave extended cognition (see Sutton 2006, 2010). If we set out to investigate individual differences in the pursuit of epistemic hygiene,

[19] See Andrada 2020 on the importance of attending to diverse embodiments when giving an account of the phenomenology of extended cognition.

the question is not so much the personality or inclinations of a particular agent (for instance, how meticulous they are), but rather how a particular embodiment contributes to fixing the type of engagement that might be required for achieving a minimal form of epistemic hygiene.

To illustrate, let us focus on the reliability of our extended and unextended cognitive processes. Our senses (hearing, sight, etc.) have a long history of evolutionary testing and, when working well, they are tuned to our environment; hence they are for the most part reliable. In this regard, reliability is the result, to a large extent, of the good working of our sub-personal cognitive architecture. By this, I mean that in most cases, we, as cognitive agents, do not have to do much for our *basic* cognitive abilities to be reliable. The problem is that this appreciation, combined with the quest for similarities prevalent in an epistemic parity approach, might mislead us into thinking that this is how reliability is accomplished in *all* instances of extended cognition and knowledge.[20]

The epistemic complementarity framework calls for a more fined-grained analysis. Recognizing the differences between extended and unextended processes, together with the vital role of embodied manipulations and skilled gestures in extended cognitive processes, makes us realize that the type of engagement required for establishing their reliability can vary from one person to another. This means that despite the fact that the kind of coarse-grained cognitive function implemented by an unextended and an extended cognitive process (e.g. *biomemory* and extended memory) can be thought of as the same, what it takes for them to be reliable might be different; more precisely, and importantly, it might differ with respect to the person-level demands it makes on the agent. This will become clearer in Sect. 12.4, where I will revisit the traditional case of Otto and his notebook (Clark and Chalmers 1998), and address the bearing that Otto's cognitive profile (i.e. his mild form of dementia) has on his pursuit of epistemic hygiene.

12.6 Step 2: Properties of the Technology

The second step of an epistemic complementarity framework takes us to an analysis of the technology's properties. The central idea here is that the very features of the technology have a bearing on the epistemic standing of the extended cognitive process. As we saw before, the central tenet behind an epistemic complementarity-based extended epistemology is that all the elements that constitute an extended cognitive process have a bearing on the pursuit of epistemic hygiene. Obviously different technologies vary drastically from each other. So we need to specify how the diversity of technological properties matters *epistemically*.

[20] See for instance Clark 2015.

When discussing the reliability of a certain technology, we can refer to two different but interrelated aspects.[21] First, a technological device can be reliable in terms of its robustness or resilience. Second, when it comes to the reliability of the information it conveys, we may refer to its factive status. Despite their differences, both of these aspects are relevant for granting a positive epistemic standing to an extended cognitive process. This is because reliability may be a function of the information such technologies support, which in some cases is not entirely disconnected from their working and performing correctly. If we refer to a technological device as being reliable, we are saying that it operates in accordance with our expectations, in terms of both resilience and reliability of content.

The relevance of taking into account the properties of the technology when identifying the individual engagement required for achieving epistemic hygiene, is that it will in fact vary depending on them. This means that when investigating the epistemic standing of a pervasive interaction with a given technology, such as smartphone, we must look not only at the individual agent and her internal cognitive character, but also at the reliability of the technology itself.

For instance, Heersmink (2018) and Heersmink and Sutton (2018) have recently proposed a virtue responsibilist approach to bolstering our epistemic credentials while interacting with the Internet. This basically means instilling virtues such as honesty, diligence and thoroughness into one's character through learning and education. Although this is certainly important, the idea behind the epistemic complementary framework is precisely that, given the distributed nature of cognition and the active nature of external elements, epistemic success does not rest solely on the individual.[22] This means that we must also look at the different epistemic properties of the technologies themselves, for instance, how data and information is recorded and stored, how easily can it be altered or modified, etc. Along these lines, Paul Smart (2018b) has highlighted the importance of taking into account the properties of internet technologies when it comes to analyzing internet-based knowledge, and not just the behavior of the individual agents who deploy or rely on such technologies. The characteristics of an online environment will affect what it takes for an agent to achieve knowledge when relying on such an environment. We can say the same of other technologies we deploy and rely on.

One implication of this analysis is that extended epistemologists should get a better grip on the properties of the actual technologies that we pervasively and stably interact with. This will help us achieve a more realistic picture concerning what it actually takes to achieve extended knowledge in our contemporary high-tech environments. An epistemic complementarity framework is thus poised to bring extended epistemology into contact with vibrant interdisciplinary debates.

[21] Thanks to an anonymous reviewer for helping me to clarify this point.

[22] Both John Sutton and Richard Heersmink are leading advocates of a complementarity-based approach to extended cognition. Inspired by their work, I have developed the epistemic complementarity principle whereby all the elements that constitute a cognitive system have a bearing on the epistemic standing of an individual's cognitive process.

12.7 Step 3: Socio-cultural Environment

Finally, the last step that needs to be taken to elucidate the demands of epistemic hygiene on extended cognitive processes concerns the socio-cultural environment in which the interaction takes place. Cognition does not take place in a vacuum, but rather is embedded in a normative environment (Menary 2007, 2012). An epistemic complementarity approach starts from the idea that elements of the social and cultural environment have an impact on the epistemic standing of an extended cognitive process, and hence are relevant for determining the type of engagement required for achieving a minimal form of epistemic hygiene.

Recognizing the role played by the socio-cultural environment involves more than simply identifying where the interaction happens; it also involves taking account of any effects that the cultural environment might have on our epistemic processes. This has also been studied in feminist epistemologies, where it is widely accepted that knowers are entangled in social relations, some aspects of which have a bearing on their epistemic life.[23] Moreover, the cultural nature of many of our interactions with technologies has been made clear by leading advocates of extended cognition, although it has not been sufficiently taken on board in extended epistemology.[24]

To illustrate this, I will address the influence that social practices have on the type of individual engagement required to be minimally hygienic epistemic agents. By 'social practices' I mean patterns of action spread over a population, which are transmitted both between the members of a community and across generations.

Let us begin by looking at the role of social practices in extended cognition. The basic idea that we need to incorporate into our extended epistemology is that our skilled interactions with technologies require the acquisition of several social practices. These practices include *cognitive* practices of a specific type, which are acquired in virtue of the process of enculturation (Menary 2007). Some examples of cognitive practices are the manipulation of public systems of symbolic representation (such as mathematics, reading and writing), and the practice of epistemic diligence concerning the structuring and maintenance of the quality of information stored in the environment (Menary 2012, 2018b). Enculturation is a form of non-genetic inheritance, and such practices are quite recent in phylogenetic terms; hence social practices and cumulative cognitive niches are needed to guide our learning histories (Menary 2018a).

Acknowledging the role that cognitive practices have in our extended cognitive processes is relevant for determining what is required for an individual agent to be epistemically hygienic, mainly because the type of engagement will vary according to such practices. For instance, cognitive practices are acquired mainly through interpersonal relations (e.g. infant and caretaker), and some hygienic practices

[23] See for instance Haraway 1988, Haslanger 2007, 2020.

[24] To the best of my knowledge, exceptions include Menary 2012, 2018b, and Kotzee 2018.

might also be institutionally supported, in the form of standards, educational programs or pedagogical methods. These reliable practices might make the type of individual engagement less demanding. However, this cuts both ways, in the sense that a lack of practices, or the presence of unreliable ones, might place more demands on the individual agent.[25]

This interplay between individual agent and social practices is connected to a recent account of our social and material epistemic dependence.[26] According to Sanford C. Goldberg (2017a), whether the members of our epistemic community do or do not comply with the relevant norms governing our social practices has a positive or negative impact on our epistemic standing.

This can be illustrated by looking at our epistemic reliance on technologies. To simplify somewhat, the idea is that if someone relies on a technological device made by a socially recognized manufacturer, they are entitled to expect that the manufacturer has complied to the social norms and standards regarding the manufacture of the device. The agent has, to use Sandy Goldberg's terminology, a practice-generated entitlement (Goldberg 2017b, 2018). Thus they are entitled to expect that the information conveyed is accurate and reliable, at least insofar as the designers comply to the norms of design and the user complies to the norms of use (e.g., follows the instruction manual and performs all relevant maintenance activities).

I take this analysis as evidence that the type of engagement that is required to be even minimally epistemically hygienic will vary depending not only on the individual's acquisition of social practices, but also on how other members of our epistemic community act. For instance, flaws exhibited by members of our epistemic community might make it harder for an individual agent to acquire epistemic goods, but also their active contribution might relieve some of their individual burden.

Importantly, this type of epistemic dependence on others is even more "profound", as Goldberg (2017a) remarks. It is manifested not only in our reliance on instruments, but also through the design of our shared environments. To elucidate one possible way in which this could happen, Goldberg appeals to the notion of 'epistemically engineered environments'. By that, Goldberg means "an environment that has been deliberately designed so as to decrease the cognitive burden on individual subjects in their attempts to acquire knowledge" (Goldberg 2017a). For instance a *classroom* is a learning environment in which students benefit from the specific design which is "pre-screened, and chosen with an eye on epistemic standards". Social practices can thus guide the design of specific shared environments,

[25] Care is needed here insofar as, given the long time spans of enculturation, the reliability of a given practice might take transgenerational intervals to be understood or even recognized. For more on this, see Levy and Alfano 2020.

[26] For a thorough taxonomy of the varieties of epistemic dependence, see Broncano-Berrocal and Vega-Encabo 2020.

and can promote and enhance the epistemic goods available to the individual agent, and minimize their cognitive load.[27]

To sum up, an epistemic complementarity approach compels us to attend to the interplay between the embodied profile of a cognitive agent, the properties of a technology and the socio-cultural environment in which the interaction takes place. All of these elements contribute to the epistemic hygiene of an extended cognitive process. In the next section, I will briefly illustrate how this model works by applying it to a case in which an agent interacts with a technological device.

12.8 Epistemic Complementarity Meets Otto

The deeply anchored individualism that characterizes most theorizing about epistemic and cognitive agency makes us more likely to recognize the active role of the social and material environment in cases that concern people with compromised organic cognitive systems. This is in fact why the most compelling cases of extended cognition, including the one I will present here, concern an agent suffering from Alzheimer's disease. However, we should not forget that the central idea of the epistemic complementarity framework is that the social and material world plays an active role in our cognitive and epistemic lives, whether or not our organic cognitive faculties are compromised.

The canonical example of extended cognition, since the publication of Clark and Chalmers's seminal paper, is the case of Otto and his notebook. Otto is someone who suffers from a mild form of Alzheimer's disease and heavily relies on a memory notebook. He writes down every new piece of information he acquires, and looks it up whenever he wants to perform an action. Given his pervasive interaction with it, the notebook is part of his (extended) memory system. Otto's interaction with his notebook is contrasted with Inga's interaction with her biological memory when undertaking an action. The case I will analyze is an adaptation of Otto's case.

Currently smartwatches are being introduced into healthcare practices, as a form of assistive technology for people with dementia and Alzheimer's (Thorpel et al. 2016). Hence I will apply the epistemic complementarity model to a revised version of Otto's case where he interacts with a smartwatch, as follows:

> **Smart Otto**: Otto suffers from a mild form of dementia. He wears a smartwatch wherever he goes, which helps him to organize his daily routines. He has developed the following habit: whenever he finds something interesting or worth remembering, he stores the corresponding information using an app which allows him to record voice notes. During the

[27] Goldberg 2017a identifies the effect of this sort of epistemic dependence by the status of epistemic justification. In this respect, the account on offer is one in which our epistemic dependence on designers and manufactures entails that their behavior with respect to the design of an instrument, or even an environment, can undermine or defeat the belief we form in virtue of using such an instrument. Here I am adopting a more general perspective, by focusing on the type of engagement required for achieving a minimal form of epistemic hygiene.

> afternoon, Otto carefully updates his information concerning future plans (and directions, in case he has to go to certain places) in a different app which supports reminders. Once a fortnight, Otto uses this information to meet his friend Inga and head to the MoMA museum.

The question that needs to be addressed is what makes Otto a minimally hygienic epistemic agent, to the extent that we can attribute knowledge to him (for instance, to the effect that he knows when he is meeting Inga, and the address of the MoMA Museum, etc.). We must take for granted that he relies on his smartwatch in quite an intimate way, and that his interaction with his smartwatch is integrated with his other cognitive routines, just like in the original case.

According to the epistemic complementarity framework, the type of engagement required for being minimally epistemic hygienic is determined by the interplay between the embodied agent, the properties of the relevant technology, and the socio-cultural environment in which they are embedded. Minimal epistemic hygiene can be understood as the reliability (generally truth-conducive) of the belief-forming process and, importantly, reacting to the shifting reliability of this process. This means, for instance, that if the process is unreliable, the agent reacts accordingly (i.e. does not trust it). We have established that this can be done either actively or passively (sub-personally).[28]

First of all, we should look at the embodied and cognitive profile of the agent. The story tells us that Otto suffers from a mild form of dementia. People with dementia experience, as part of their progressive cognitive impairment, short-term memory problems, including language deficits, difficulties in initiating tasks, planning, monitoring and regulating behavior, visuospatial difficulties, agnosia and apraxia.[29] Accordingly, interacting with a smartwatch and engaging in different epistemic practices is not an easy task for them.

Second, we must attend to the properties of smartwatches. A smartwatch is portable and attached to the agent; consequently it is less likely that an individual with dementia would forget about it. This means that it makes less stringent cognitive demands on agents who use it, at least as far as reliability *qua* consistency and robustness is concerned. Like smartphones, these devices support a wide variety of apps, since they are run by similar operating systems. The reliability of its information partly depends on the correct functioning of the apps it supports, and partly on the reliability of the user's implemented content, since smartwatches are open to a fairly high degree of customization.

Third, we must attend to the socio-cultural environment in which the interaction takes place. First of all, we should look at the social practices surrounding Otto's interaction with his smartwatch. In order for such interaction to be reliable, Otto needs to learn, train and acquire complex patterns of action. In fact, if we look at studies of real patients who compensate for their ill-functioning biological memory

[28] Remember that this is a schematic characterization, and that in real life we might deploy both strategies. See Andrada 2019 and Andrada 2020 for more on the role of conscious epistemic care and extended cognition.

[29] World Health Organization: *Towards a dementia plan: a WHO guide* http://www.who.int/mentalhealth/neurology/dementia/en/. Last accessed: August 31, 2018.

by successfully learning to deploy smartwatches, we see that their effectiveness relies on careful training (Boletsis et al. 2015). That is why, the reliability of Otto's extended cognitive process is largely a matter of reliable practices.

Let us imagine for the sake of the argument, that Otto relies on his smartwatch but has not been trained to do so; thus he follows no method of organization. Moreover, Otto lives in a place where there are no caregiving practices that can help him acquire the relevant hygienic practices. In this situation, relying on the smartwatch is a highly demanding process for him. However, his position could be enhanced by instilling practices that help him structure the information diligently; and with the proper scaffolding, Otto could be trained to do so to the extent that he might end up being much better off. The fact that many of these practices are social in nature lends further support to the idea that the reliability of extended cognitive processes requires more than the good working of Otto's sub-personal mechanisms. Without such a display, not only might Otto struggle more, but he might also be unable to acquire knowledge in virtue of his interaction with the smartwatch. This shows why the engagement required from Otto will vary according to such practices or any lack thereof.[30]

Another aspect of the socio-cultural environment's contribution to Otto's interaction with his smartwatch can be illustrated by pointing out that Otto is entitled to expect that certain epistemic norms (e.g. norms concerning minimal reliability and GPS accuracy) were followed by the people who designed his smartwatch. This compliance might take some weight off his shoulders, to the extent that he is entitled to trust its output without actively inspecting its reliability.

However, reacting appropriately to any shifts in the reliability of this belief-forming process might be harder to do, depending on the stage of Otto's progressive cognitive impairment. If someone tampers with the information in Otto's smartwatch, it might be very hard for Otto to detect those changes and act accordingly. In fact, according to ethnographic work done with people with dementia, such individuals are heavily dependent on technologies and caregivers, usually family members, in order to conduct their daily lives (Boletsis et al. 2015; Yatczak 2018). That is why we should be open to the idea that some degree of monitoring or maintenance of Otto's belief-forming process might be distributed to the technology (for instance, in the form of monitoring apps that detect when things are not working normally and let Otto know through buzzes and vibrations), as well as to the specific layout of the environment he inhabits, and also, importantly, to other member's of

[30] I want to remark that many of our cognitive abilities are *enculturated*, although they might not be extended in the sense that concerns me here; that is, their material realizers might not be partly constituted by something external to the organism. This might be the case, as I have previously stated, for basic abilities and tasks, but not in many human cognitive activities. This should lead us to revisit the idea that reliability in intracranial cognition is entirely sub-personally achieved.

Otto's epistemic community.[31] Without them, Otto's epistemic and cognitive life might be severely compromised.[32]

All this suggests that, in order for Otto's epistemic hygiene to be less individually demanding (or even feasible), we might need to revise our previous formulation of the case. This could leave us with the following case:

> **Smart Otto (and Greg)**. Otto suffers from a mild form of dementia. He wears a smartwatch wherever he goes, which helps him to organize his daily routines. He has developed the following habit: whenever he finds something interesting or worth remembering, he stores the corresponding information using an app which allows him to record voice notes. During the afternoon, Otto and his caregiver Greg carefully update Otto's information concerning future plans (and directions, in case they have to go to certain places) in a different app which supports reminders (including buzzes). Once a fortnight, Otto uses this information to meet his friend Inga and head to the MoMA museum.

This case more accurately resembles a real-life setting. We should note that if we had followed an epistemic parity approach, we would not have been able to properly understand the many factors that contribute to Otto's epistemic life, or the engagement required for him to acquire epistemic goods. On the contrary, following the steps set out by an epistemic complementarity framework provides us with a more comprehensive picture. However, one might want to object right away that a *veridic* case such as that of *Smart Otto (and Greg)* is not an instance of extended cognition, nor of extended knowledge, precisely because of a lack of individual control, due to the heavy dependence on artifacts and other people (see Drayson and Clark, *forthcoming*, p. 24). The epistemic complementarity framework, invites us to reflect on the active role that social, material and cultural factors play in our (extended) epistemic lives. That is why figuring out the distribution of this dependence, where the relevant individual might not be the principal locus of control, is one of the central challenges for an epistemic complementarity approach to extended epistemology.

12.9 Conclusion

Let me recapitulate the main outcomes of this chapter. Its central aim was to propose a framework for modeling extended epistemology, based on second-wave extended cognition. To this end, first I have shown how modeling cognitive extension from a first-wave extended approach favors an epistemic parity principle, according to which the type of individual engagement required to ensure that an

[31] We can imagine an app that warns Otto (for instance through a vibration) of a failure in performance, or alerts his caregivers. Currently there are many apps that support complex ways of self-tracking and monitoring, many of which are deployed by people with dementia and their caregivers. See Lindqvist et al. 2015.

[32] The extent to which such epistemic monitoring might be completely outsourced to technologies is an empirical question, which raises complex ethical and sociological challenges. Notice also that this sort of extended monitoring might be used for cognitive and epistemic enhancement (see Clowes 2014).

unextended process is minimally hygienic also ensures the epistemic hygiene of an extended cognitive process. In contrast with this approach, I have introduced an epistemic complementarity principle, drawing from second-wave extended cognition. Its central tenet is that, given the complementary roles played by intracranial and external elements in extended cognitive processes, we cannot determine the sort of individual engagement required for them to be minimally epistemically hygienic by focusing only on what we know about intracranial cognitive mechanisms. We should not forget that their complementary roles are in fact afforded by their differences, some of which might be epistemic. This has been implemented via a threefold model, where the individual engagement required for epistemic hygiene varies according to the particularities of the embodied agent, the properties of the technology, and the social and cultural environment in which the interaction takes place.

The resulting extended epistemology is a more inclusive one, in that it vindicates the fact that not all knowers and technologies are alike and their differences, including those of their socio-cultural environments, matter epistemically. There is, of course, more conceptual and empirical work left to be done; here we have taken the first steps.

Acknowledgement For comments and discussion, I would like to thank Jesús Vega, Fernando Broncano, Richard Menary, Manuel de Pinedo, Robert W. Clowes, three anonymous reviewers, and audiences at Universidad Carlos III and Universidad Autónoma de Madrid. This work was funded by research grant 'Intellectual autonomy in environments of epistemic dependence' (FFI2017- 87395-P, MINECO/FEDER, EU), and by Portuguese national funds through FCT – Fundação para a Ciência e a Tecnologia under the project UIDB/FIL/00183/2021.

References

Andrada, G. (2019). Mind the notebook. *Synthese*. https://doi.org/10.1007/s11229-019-02365-9.

Andrada, G. (2020). Transparency and the phenomenology of extended Cognition. *Limite: Interdisciplinary Journal of Philosophy & Psychology, 15* (20), 1–17.

Antony, L. M. (2002). Embodiment and epistemology. In P. K. Moser (Ed.), *The Oxford handbook of epistemology Oxford*. Oxford University Press.

Boletsis, B., McCallum, S., & Landmark, B. F. (2015). The use of smartwatches for health monitoring in home-based dementia care. In J. Zhou & G. Salvendy (Eds.), *ITAP 2015, Part II, LNCS 9194* (pp. 15–26).

Broncano-Berrocal, F., & Vega-Encabo, J. (2020). A taxonomy of types of epistemic dependence: Introduction to the synthese special issue on epistemic dependence. *Synthese*. https://doi.org/10.1007/s11229-019-02233-6.

Carter, J. A. (2013). Extended cognition and epistemic luck. *Synthese, 190*(18), 4201–4214.

Carter, J. A. (2017). Virtue epistemology and extended cognition. In H. Battaly (Ed.), *Routledge handbook of virtue epistemology*. Routledge.

Carter, J. A., Kallestrup, J., Palermos, S. O., & Pritchard, D. (2014). Varieties of externalism. *Philosophical Issues, 24*(1), 63–109.

Carter, J. A., Clark, A., Kallestrup, J., Palermos, S. O., & Pritchard, D. (Eds.). (2018a). *Extended epistemology*. Oxford: Oxford University Press.

Carter, J. A., Clark, A., & Palermos, S. O. (2018b). New humans, ethics, trust. In *Extended epistemology*. Oxford University Press.

Cash, M. (2013). Cognition without borders: "third wave" socially distributed cognition and relational autonomy. *Cognitive Systems Research, 25*, 61–71.
Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. Oxford University Press.
Clark, A. (2010) "Coupling, Constitution and the Cognitive Kind: A Reply to Adams and Aizawa", in Menary, R. (Ed.), The Extended Mind, The Extended Mind, Cambridge: The MIT Press, 81–99
Clark, A. (2015). What 'extended me' knows. *Synthese, 192*(11), 3757–3775.
Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*, 1.
Clowes, R. (2014). Thinking in the cloud: The cognitive incorporation of cloud-based technology. *Philosophy and Technology, 28*(2), 261–296.
Clowes, R. (2018). Immaterial engagement: Human agency and the cognitive ecology of the internet. *Phenomenology and the Cognitive Sciences*, 1–21.
Donald, M. (1991). *Origins of the modern mind: Three stages in the evolution of culture and cognition*. Harvard University Press.
Fenton, A., & Krahn, T. (2007). Autism, neurodiversity, and equality beyond the "normal". *Journal of Ethics in Mental Health, 2*(2), 2.
Gallagher, S. (2013). The socially extended mind. *Cognitive Systems Research, 25–26*, 4–12.
Goldberg, S. (2017a). Epistemically engineered environments. *Synthese*, 1–20.
Goldberg, S. (2017b). Should have known. *Synthese, 194*(8), 2863–2894.
Goldberg, S. C. (2018). *To the best of our knowledge: Social expectations and epistemic normativity*. Oxford University Press.
Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies, 14*(3), 575–599.
Haslanger, Sally (2007). "But mom, crop-tops are cute!" Social knowledge, social structure and ideology critique. Philosophical Issues 17 (1):70–91.
Haslanger, S. (2020). Cognition as a social skill. *Australasian Philosophical Review*.
Heersmink, R. (2014). Dimensions of integration in embedded and extended cognitive systems. *Phenomenology and the Cognitive Sciences, 13*(3), 577–598.
Heersmink, R. (2018). A virtue epistemology of the internet: Search engines, intellectual virtues, and education. *Social Epistemology, 32*(1), 1–12.
Heersmink, R., & Sutton, J. (2018). Cognition and the web: Extended, transactive, or scaffolded? *Erkenntnis*, 1–26.
Kirchhoff, M. D. (2012). Extended cognition and fixed properties: Steps to a third-wave version of extended cognition. *Phenomenology and the Cognitive Sciences, 11*(2), 287–308.
Kirchhoff, M. D., & Kiverstein, J. (2019). Extended consciousness and predictive processing: A third-wave view. London, UK: Routledge.
Kiverstein, J., & Farina, M. (2011). Embraining culture: Leaky minds and spongy brains. *Teorema, 32*(2), 35–53.
Kotzee, B. (2018). Cyborgs, knowledge and credit from learning. In J. A. Carter, A. Clark, J. Kallestrup, S. O. Palermos, & D. Pritchard (Eds.), *Extended epistemology*. Oxford: Oxford University Press.
Levy, N., & Alfano, M. (2020). Knowledge from vice: Deeply social epistemology. *Mind, 129*(515), 887–915.
Lindqvist, E., Larsson, T. J., & Borell, L. (2015). Experienced usability of assistive technology for cognitive support with respect to user goals. *NeuroRehabilitation, 36*, 135–149.
Menary, R. (2006). Attacking the bounds of cognition. *Philosophical Psychology, 19*(3), 329–344.
Menary, R. (2007). *Cognitive integration: Mind and cognition unbounded*. New York: Palgrave Macmillian.
Menary, R. (2010). Cognitive integration and the extended mind. In R. Menary (Ed.), *The extended mind*. Cambridge: MIT Press.
Menary, R. (2012). Cognitive practices and cognitive character. *Philosophical Explorations, 15*(2), 147–164.

Menary, R. (2018a). Cognitive integration. How culture transforms us and extends our cognitive capabilities. In S. Gallagher, A. Newen, & L. De Bruin (Eds.), *Oxford handbook of 4E cognition* (pp. 187–215). Oxford: Oxford University Press.
Menary, R. (2018b). Keeping track with things. In J. A. Carter, A. Clark, J. Kallestrup, S. O. Palermos, & D. Pritchard (Eds.), *Extended epistemology* (pp. 305–330).
Nicolle, L. (2007). Hygiene: What and why. *Canadian Medical Association Journal, 176*(6), 767–768.
Nisbett, R., Choi, I., Peng, K., & Norenzayan, A. (2001). Culture and Systems of Thought: Holistic vs. analytic cognition. *Psychological Review, 108*(2), 291–310.
Palermos, O. S. (2014). Knowledge and cognitive integration. *Synthese, 191*, 1931–1951.
Pritchard, D. (2010). Cognitive ability and the extended cognition thesis. *Synthese, 175*, 133–151.
Pritchard, D. (2018). Extended epistemology. In A. Carter, A. Clark, J. Kallestrup, O. S. Palermos, & D. Pritchard (Eds.), *Extended epistemology* (pp. 90–104). Oxford: Oxford University Press.
Rowlands, M. (2009). The extended mind. *Zygon, 44*, 628–641.
Rowlands, M. (2010). *The new science of the mind: From extended mind to embodied phenomenology*. Bradford.
Smart, P. R. (2018a). Emerging digital technologies: Implications for extended conceptions of cognition and knowledge. In A. J. Carter, A. Clark, J. Kallestrup, O. S. Palermos, & D. Pritchard (Eds.), *Extended epistemology* (pp. 266–304). Oxford: Oxford University Press.
Smart, P. R. (2018b). (Fake?) news alert: Intellectual virtues required for online knowledge! *Social Epistemology Review and Reply Collective, 7*(2), 45–55.
Sprevak, M. (2009). Extended cognition and functionalism. *Journal of Philosophy, 106*(9), 503–527.
Sterelny, K. (2010). Minds: Extended or scaffolded? *Phenomenology and the Cognitive Sciences, 9*, 465–481.
Sutton, J. (2006). Distributed cognition: Domains and dimensions. *Pragmatics and Cognition, 14*(2), 235–247.
Sutton, J. (2010). Exograms and interdisciplinarity: History, the extended mind and the civilizing process. In R. Menary (Ed.), *The extended mind* (pp. 189–225). Cambridge, MA: MIT Press.
Sutton, J., Harris, C. B., Keil, P. G., & Barnier, A. J. (2010). The psychology of memory, extended cognition, and socially distributed remembering. *Phenomenology and the Cognitive Sciences, 9*(4), 521–560.
Thorpel, J. R., Rønn-Andersen, K. V. H., Bien, P., Gurcan Ozkil, A., Forchhammer, B. H., & Maier, A. M. (2016). Pervasive assistive technology for people with dementia: A UCD case. *Healthcare Technology Letters, 4*, 297–302.
Wadham, J. (2015). Common-sense functionalism and the extended mind. *The Philosophical Quarterly, 66*(262), 136–151.
Yatczak, J. (2018). Everyday material engagement: Supporting self and personhood in people with Alzheimer's disease. *Phenomenology and the Cognitive Sciences*, 1–18.

Gloria Andrada is a research fellow at the Lisbon Mind and Reasoning Group, Nova University of Lisbon, and a PhD candidate at the Autonomous University of Madrid. She works at the intersection of epistemology and philosophy of the cognitive sciences. Her current research focuses primarily on extended cognition, extended knowledge, and on the bearing of emerging digital technologies on our cognitive abilities and epistemic practices.

Chapter 13
The Extended Mind: A Chapter in the History of Transhumanism

Georg Theiner

13.1 Introduction

To set the stage for my reconstruction, I begin by reviewing a few central tenets of transhumanist thinking, and position them vis-à-vis the EMT. Transhumanism is an intellectual, cultural, and political movement promoting the development and use of advanced technologies to improve human capacities, and to enhance human lives beyond their current form.[1] James Hughes (2004), a proponent of this viewpoint, asserts that transhumanists generally subscribe to the thesis that "human beings should use technology to transcend the limitations of the body and brain" (Hughes 2004, p. 155). Nick Bostrom (2005), another champion of transhumanism, defines the movement as "an outgrowth of secular humanism and the Enlightenment" which "holds that current human nature is improvable through the use of applied science and other rational methods, which may make it possible to increase human health span, extend our intellectual and physical capacities, and give us increased control over our own mental states and moods" (Bostrom 2005, pp. 202–203). The history of the term "transhumanism", in its contemporary usage, goes back to an influential essay by the British evolutionary biologist and first Director of the UNESCO, Julian

Zwei Seelen wohnen, ach! in meiner Brust. (Goethe, Faust)

[1] For an introduction, see, e.g., Hughes (2004), the articles collected in More and More (2013), Ranisch and Sorgner (2014), and MacFarland (2020). For an overview of key topics, see the "Transhumanist FAQ" (version 3.0), available at https://humanityplus.org/philosophy/transhumanist-faq/. The most up-to-date repository of information on various strands of transhumanist thinking is the wiki-style *HPlus Pedia* (2019), available at https://hpluspedia.org/wiki/Main_Page

G. Theiner (✉)
Department of Philosophy, Villanova University, Villanova, PA, USA
e-mail: georg.theiner@villanova.edu

R. W. Clowes et al. (eds.), *The Mind-Technology Problem*, Studies in Brain and Mind 18, https://doi.org/10.1007/978-3-030-72644-7_13

Huxley (1957), who advocated the use of genetic engineering to consciously redirect the course of human evolution.[2] But going beyond Huxley's focus on biotechnological enhancement, the most recent version of the "Transhumanist FAQ" (Transhumanist FAQ, 2019) cites as inspiration the wider conception of transhumanism by Max More (1996), as "a class of philosophies of life that seek the continuation and acceleration of the evolution of intelligent life beyond its currently human form and human limitations by means of science and technology, guided by life-promoting principles and values".

One reason for endorsing More's more inclusive conception is that the cumulative effects of the envisioned technological interventions to the human body and mind are bound to produce beings that are sufficiently distinct from our shared biological heritage so as to warrant a description as "posthuman". Understood in this context, the current use of "transhuman" as a shorthand for "transitional human" was popularized by the Iranian-American futurist thinker Fereidoun M. Esfandiary (1989), better known as FM-2030. The transition from human to posthuman might follow several different paths, which can be seen as representing different strands of transhumanist thinking.[3] Here, I only mention two divergent varieties. Representing the biology-centered wing, Aubrey de Grey (de Grey and Rae 2007) and his colleagues are plotting to engineer "negligible senescence" in humans, by developing "disruptive" biomedical technologies that would stop or even reverse the effects of aging, age-related cellular and molecular defects, with the ultimate goal of achieving biological immortality. Representing the cybernetics-centered wing, the inventor and AI pioneer Ray Kurzweil (1999, 2005) yearns for a future in which our posthuman descendants will have shirked their carbon-based biological bodies, having attained the ability to upload digitized versions of themselves onto computers. We shall return to the intellectual roots of these two competing projections of humanity below.

In my paper, I mostly deal with two comparatively "modest" transhumanist aspirations, which – following the six-fold taxonomy developed by Fuller (2011, pp. 103–05) – fall under the categories of "Humanity Enhanced" and "Humanity Incorporated". The first category involves strategies and technologies for improving various traits or capacities that human beings already possess, at least in a rudimentary manner, but deployed with the goal of elevating them well beyond current levels (Savulescu and Bostrom 2009). Those include both physical and mental aspects of human existence. An example of physical enhancement would be the use of nanotechnology to reduce the deposit of fatty arteries; an example of cognitive

[2] Contrary to popular belief, Harrison and Wolyniak (2015) point out that Huxley did not *coin* the term. He first used it in a lecture published in 1951 (and not in his 1927 essay), but the term itself has a long history that dates back to Dante's *Paradiso* and the Pauline epistles. For a defense of genetic enhancement from a utilitarian standpoint, without necessarily taking on board the "transhumanist" agenda, see Harrison (2007).

[3] See, for example, the taxonomy of "futurist ideas and positions" available at https://hpluspedia.org/wiki/Comparison

enhancement would be the consumption of mind-altering drugs, the installment of neural implants, or the use of brain-machine interfaces. More generally, Bostrom and Sandberg (2009) define cognitive enhancements broadly as "the amplification or extension of core capacities of the mind through improvement or augmentation of internal or external information processing systems" (Bostrom and Sandberg 2009, p. 311). As they point out, cognitive enhancements in this wider sense, achieved, e.g., through dietary and educational regimes, cultural norms, moral edification, ascetic practices, or legislative efforts (cf. Bostrom and Sandberg 2009) have long been a hallmark feature of humanity. However, the morally controversial application of "Humanity Enhanced" concerns the strategic application of scientifically advanced technologies to boost human performance. Still, the above definition is meant to include interventions contrived to induce subjective experiences that would traditionally be called spiritual or mystical. To name but a few, those would include strapping on a "god helmet" (Persinger 1991) – i.e., a head-mounted apparatus for generating low-intensity magnetic fields that purportedly elicit visions of a divine presence; or the use of artificial "entheogens" (Ruck et al. 1979), i.e., psychoactive substances specifically designed and consumed to promote spiritually significant experiences in its users. Relatedly, I also note here that Bostrom and Sandberg's generic use of the term "cognitive" is meant to include things like affect, emotions, moods, and psychological well-being, which transhumanists equitably take to be potential targets of enhancement (Hughes 2004; Bostrom 2008). However, since Clark's EMT is concerned mostly with intellective modes of cognition, affective and other psychological dimensions of human enhancement will similarly take a back seat in my own discussion.[4]

With his category of "Humanity Incorporated," Fuller (2011) refers to the distinctively human proclivity for continually negotiating the boundary conditions of our organic mode of existence - either by making environmental resources literally part of ourselves, or by combining human and non-human elements into emergent wholes that yield novel, otherwise unattainable capabilities (Clark 2003, 2008). Historically speaking, it would be hard to exaggerate the profound impact of the human drive towards incorporation. In the twentieth century, perhaps the most visionary treatment of this theme can be found in the work of Marshall McLuhan, the Canadian-born father of media theory and oft-cited prophet of the information age (Logan 2010; Sandstrom 2014). In *Understanding Media: Extensions of Man* (1964), McLuhan preceded Clark's EMT with his suggestion to view tools and media literally as extensions of the human body and mind. In particular, McLuhan anticipated that a dramatic shift would be taking place as "electric" information and communication technologies were beginning to supplement our much older

[4] But see, e.g., Robinson (2013), Greenwood (2015), and Krueger and Szanto (2016) for extended mind-friendly treatments of affect and emotion.

reliance on "mechanical" technologies as replacements or extensions of human body parts[5]:

> "In this electric age we see ourselves being translated more and more into the form of information, moving toward the technological extension of consciousness. […] By putting our physical bodies inside our extended nervous systems, by means of electric media, we set up a dynamic by which all previous technologies that are mere extensions of hands and feet and bodily heat-controls - all such extensions of our bodies, including cities - will be translated into information systems." (McLuhan 1964, p. 20)

The spiritual dimension of McLuhan's conception of media can be gleaned from the fact that among his key inspirations was the thought of the nineteenth century transcendentalist Ralph Waldo Emerson, who praised the "mechanical contrivances" of his day and age thusly:

> "The human body is the magazine of inventions, the patent office, where are the models from which every hint was taken. All the tools and engines on earth are only extensions of its limbs and senses. One definition of man is 'an intelligence served by organs.' Machines can only second, not supply, his unaided senses." (Emerson 2008, p. 158)

A systematic emphasis on the transformative effects of incorporating natural resources into systems of human value, through the process of labor, can be found in the writings of Marx and Engels. With an homage to Darwin, whom Marx held in high esteem for his discovery that the (re-) productive relationships between organisms and their environment are in perpetual flux, and thus subject to historical change, Marx avers in a footnote of *Capital* that

> "Technology discloses man's mode of dealing with Nature, the process of production by which he sustains his life, and thereby also lays bare the mode of formation of his social relations, and of the mental conceptions that flow from them." (Marx 1906, p. 406, n. 2).

In post-Revolutionary Russia, Lev Vygotsky, Alexander Luria and other members of the "Vygotsky Circle" took up the task of shaping a broadly Marxian anthropology into a distinctively "Soviet" brand of scientific psychology (Kozulin 1984). Consonant with the dialectical materialism of Marx and Engels, the Vygotskians upheld the distinction between human praxis and animal behavior by stressing the voluntary and purposive nature of the former. This led them to oppose simplistic reductions of mind to reflex-like behavior and purely mechanical processes, during an era in which behaviorism was a dominant force on the Russian psychological scene. At the same time, their dialectical-materialist heritage taught them to avoid the metaphysical trappings of Cartesian mind-body dualism, for its tendency to drive a deep wedge between an "inner" world of pure consciousness and an "outer" world of social and material reality. Thus, while trying to steer clear of both the Scylla of behaviorism and the Charybdis of dualism, the Vygotskians founded their Marxist psychology on a notion of *activity* as a culturally and historically situated, and socially and materially mediated process (Kozulin 1984). Their approach subsequently spurred the cross-disciplinary framework of Cultural-Historical Activity

[5] Brey (2000) gives an excellent survey of technologies that act as extensions of the human body.

Theory (CHAT) in the West, developed most notably by Michael Cole (1996) and Yrjö Engeström (1993) as an alternative to the reigning "intellectualism" within mainstream cognitive psychology.

Central to the Vygotskian conception of *activity* is the special emphasis on the action-mediating role of tools and symbols, which is considered as transformative of the productive relationships between subject and object, and between individual and community (Theiner and Drain 2017). The Vygotskians thus clearly recognized the evolutionary role of tool use for the culturally mediated emergence of (what they called) the "higher" from more "elementary" psychological functions: "The central characteristic of elementary functions is that they are totally and directly determined by stimulation from the environment. For higher functions, the central feature is self-generated stimulation, that is, the creation and use of artificial stimuli which become the immediate causes of behavior" (Vygotsky 1978, p. 39). Highlighting the momentous role of human labor and tool use – itself contingent on the historical development of society – in scaffolding the higher psychological functions, Vygotsky and Luria specifically stress how "[p]erfecting the 'means of labor' and the 'means of behavior' in the form of language and other sign systems, that is, auxiliary tools in the process of mastering behavior, occupies first place, superseding the development of 'the bare hand and the intellect of its own'[6]" (Vygotsky and Luria 1993, p. 78). In their assessment of the net effect of technological augmentation, they note that "[c]ultural man does not have to strain his vision to see a distant object – he can do it with the help of eyeglasses, binoculars, or a telescope; he does not have to lend an attentive ear to a distant source, run for his life to bring news, – he performs all these functions with the help of those tools and means of communication and transportation that fulfill his will. All the artificial tools, the entire cultural environment, serve to 'expand our senses'" (Vygotsky and Luria 1993, p. 169).

Don Norman, a guru of artifact design whose pioneering analyses of human-computer interaction contributed to Apple's development of the first graphical user interface in the early 1980s, drew heavily on Vygotsky's ideas on tool-based cognitive mediation (albeit filtered through the lens of his colleague at UCSD, Michael Cole, who co-edited the English 1978 edition of Vygotsky's *Mind in Society*). Norman coined the term *cognitive artifact,* defined as "those artificial devices that maintain, display, or operate upon information in order to serve a representational function and that affect human cognitive performance" (Norman 1991, p. 17). He saw that the key to designing user-friendly cognitive artifacts is to understand how the tool itself - its affordances, capabilities, and material constraints – transforms the nature of the task which a user has to perform, in a given task environment (e.g., to get a computer to do useful things by clicking on GUIs, rather than by typing arbitrary symbol strings into a DOS command shell). More generally, the guiding principle behind cognitive ergonomics is to avoid overtaxing our resource-limited

[6] With this phrase, Vygotsky refers back to a quote by Francis Bacon, which he chose as an epigraph for his book: "The bare hand and the intellect left to its own disposal is not worth very much: everything is accomplished with the help of tools and auxiliary means". In Section 3, we shall return to the Baconian undertow of EMT.

biological brains by finding ingenious ways of *distributing cognition* in space, time, and across groups of people (Norman 1991; Hutchins 1995; Hollan et al. 2000). The resultant literature on how to achieve this feat hugely influenced Clark's thoughts on extended cognition.

Viewed from Fuller's transhumanist vantage point, the EMT thus comes into view as a shrewd blend between "Humanity Enhanced" and "Humanity Incorporated" but with the added plot twist that deep down, our human minds themselves are tools or artifacts.[7] With the portrayal of human beings as *natural-born cyborgs* (Clark 2003), Clark's EMT takes up the familiar theme of "man the tool-maker" (*homo faber*), but pushes it further in two metaphysically crucial respects.

First, by considering us as *cyborgs*, the EMT blurs the customary boundary between biology and technology, or between user and tool, by asserting that extra-organismic, abiotic resources are literally part the extended machinery which constitutes our minds and selves. Clark's EMT thus effectively places *homo sapiens* into the category of human-technology symbionts, befitting Fuller's description of "Humanity Incorporated" as "an especially materialistic take on the *imago dei* doctrine" (Fuller 2011, p. 104). Fuller's reference to a key Biblical notion – common to the Abrahamic religions – that human beings have been created "in the image and likeness" of God (*imago dei*) may seem bewildering to the unsuspecting reader. Indeed, it will take the entirety of section 2 to provide the necessary historical (Christian) background that will later allow us, in sections 3 and 4, to trace out more fully a historical arc connecting it to Clark's EMT. For the moment, though, Fuller's qualification that we are dealing with a "materialistic" inflection of *imago dei* should strike us as equally if not more surprising. As we shall discuss later, traditional renderings of the putative resemblance between humans and God typically involve comparisons along intellectual, moral, or spiritual dimensions, all of which might be thought to presuppose some form of mind-body dualism (or at least hylomorphism). At the same time, positing an immaterial soul as the locus of genuine human agency, thinking, and consciousness is clearly anathema for the materialist orthodoxy of the EMT, which abhors the dualist folly of a "Cartesian theatre" (Dennett 1992) lodged somewhere inside the skull. As Clark vehemently insists, "there is no single, all-powerful, hidden agent inside the brain whose job is to do *all the real thinking* and which is able to intelligently organize all those teams of internal and external supporting structure. Indeed, on the most radical model that we have scouted, it is (as it were) supporting structure 'all the way down,' with mind and reason the emergent products of a well-functioning swirl of (mostly) self-organizing complexity" (Clark 2008, p. 136).

But the second aspect of Clark's icon of the "natural-born cyborg" is at least equally intriguing, because it places our necessary *openness* to information-processing mergers and coalitions at the heart of what it means to be human. As *natural-born* cyborgs, and thus, for Clark, quite unlike non-human animals, we are

[7] See, e.g., Heersmink (2017) for a discussion of cognitive enhancement and the extended mind thesis, and Cabrera (2015) for a discussion of different human enhancement paradigms.

obligatory human-technology symbionts. Our natural proclivity to enter into deep and profoundly transformative cognitive entanglements with tools and artifacts reveals the true "cyborg nature" of humanity (Clark 2003, p. 198). In Clark's guardedly optimistic assessment, our necessary technology dependence should not be seen as a liability but rather turns out to be humanity's greatest asset (Clark 2003, pp. 138–42). For Clark, what underwrites human uniqueness is "our special character, as human beings, to be forever driven to create, co-opt, annex, and exploit non-biological props and scaffoldings. We have been designed, by Mother Nature, to exploit deep neural plasticity in order to become one with our best and most reliable tools. Minds like ours were made for mergers. Tools-R-Us, and always have been" (Clark 2003, pp. 6–7). "Plasticity and multiplicity are our true constants," professes Clark (2003, p. 8). Elsewhere, Clark thus summarizes an important upshot of his EMT as follows: "The symbiosis of brain and cognitive technology repeated again and again but with new technologies sculpting brains in different ways, may be the origin of a golden loop, a virtuous spiral of brain/culture influence that allows human minds to go where no animal minds have gone before" (Clark 2013, p. 180). We alone, one might also say, are the self-transcending species – a leitmotif that traditionally would have been expressed in a dualist register, but which Clark resoundingly reworked into his unique brand of technologically-enhanced materialism. As he is well aware (Clark 2003, p. 139), a recognition of our deeply and inevitably cyborgian nature is meant to defuse bio-conservative calls to "protect" the "real nature" of our unfettered cognitive endowment from the "alienating" effects of intrusive technological intervention. From Clark's cyborgian perspective, most if not all of what travels under the banner of "advanced human cognition" must be seen as "always already" technologically enhanced rather than as accomplishments of our "naked," i.e., unaugmented biological brains. In this respect, Clark's EMT supplies a forceful evolutionary rationale for Fuller's theologically inspired brand of "Humanity Enhanced." Drawing an explicit analogy between the future-oriented agenda of using science and technology to enhance the human condition, and our inborn appetite for using mind-expanding tricks and tools, the EMT thus effectively "normalizes" transhumanist aspirations as the continuation of something we have been doing all along (Clark 2007).

Let us pause here for a moment, to ponder the place which the peculiar brand of Clark's naturalism, with its unique blend of both materialist and dualist motives, occupies within the larger edifice of Western philosophy. In particular, we must ask what role technology – and more specifically, technological transformation – plays for Clark's metaphysics of the human. Furthering a dualist stance, the increasing sophistication of the machines we create, which are at full display in human endeavors such as AI, space exploration, or genetic engineering, has historically often been taken as indicative of our God-like knowledge and powers, unparalleled among other species (Noble 1997). At the same time, Clark's insistence on the deep and intimate material dependence of our decentralized "soft selves" (Clark 2003) on the contributions of embodiment, culture, and environment makes us creatures of the world, which seems to offer little prospect for human aspirations to transcend the realm of materiality. Prefigured by this alternative, we can observe two

diametrically opposed metaphysical attitudes towards the human that have been defended under the equivocal, and hence potentially misleading umbrella term "posthuman" (Ferrando 2013; Ranisch and Sorgner 2014; Nayar 2014). Interestingly, both attitudes share the presumption that the human condition is essentially non-fixed and mutable, and that technology has an ontologically constitutive role to play in sculpting and transforming human nature. However, the two standpoints are fundamentally at odds with each other regarding their assessment of human exceptionalism.

On one hand, cultural posthumanists resolutely reject the idea that human beings enjoy a privileged normative standing, with the implication that they are entitled to, and also held responsible for, exerting dominion over non-human beings and Nature (writ large). In fact, cultural posthumanists loathe the metaphysically loaded category of the "human" as being hopelessly tangled up with implicitly supremacist assumptions that have historically been used as a pretext, by privileged elites, to justify the marginalization and oppression of second-grade or "less-than-fully-human" beings (Braidotti 2013). A founding essay of this tradition is Dona Haraway's "Cyborg Manifesto" (1991). Endorsing a similar line of thought, Mazlish (1993) celebrates the obliteration of the supposed "fourth discontinuity" between humans and their machines as a salutary experience, following the series of blows that our previously hyper-inflated human egos were rightfully dealt by Copernicus, Darwin, and Freud. In contrast, transhumanists continue to hold in high esteem the core ideals of the Enlightenment, which they cite in support of the proactive use of science and technology to emancipate the human condition from its evolved biological form – a process they regard as essential for realizing humanity's full potential. In short, one could say that whereas cultural posthumanism is part of the Counter-Enlightenment tradition, with its radical denial of human transcendence, transhumanism is a super-charged version of the Enlightenment which remains fully committed to the quest for human transcendence (Fuller 2011, 2017).

With his heterodox Christian development of transhumanism, Fuller is clearly heir to the Enlightenment tradition, as expressed by his conviction that "the original motivation for the West's Scientific Revolution – the radical version of Christian self-empowerment championed by the Protestant Reformation – remains the best starting point for motivating the contemporary transhumanist project" (Fuller and Lipinska 2014, p. 45). On the face of it, it would appear that Clark's EMT, developed within the broadly materialistic framework of contemporary cognitive science, has very little in common with Fuller's brazen endorsement of human exceptionalism; if anything, Clark's fondness for entering "mergers" with the endless cascades of designer environments with which we surround ourselves might suggest an affinity with the cultural posthumanist tradition. But in what follows, I set out to reverse this initial impression. That is, instead of abiding by the terms of the question Clark once posed to himself, "Post-human, Moi?" (Clark 2003, p. 197), I suggest we turn the tables on him, and ask: "Trans-human, Toi?" My hermeneutic inquiry into Clark's view takes the form of an "archeological" reconstruction of the theological backdrop to certain philosophical themes and tropes whose distant intellectual offshoots have equally informed the EMT. In particular, I aim to show that the

theological moorings of Fuller's transhumanism rest on principles that are also at play, in however tacit and materialistically inflected form, in Clark's "naturalistic" development of the EMT.[8]

13.2 Living in "Imago Dei": The Christian Foundation of Fuller's Transhumanism

Fuller's heterodox Christian vision of transhumanism is grounded in four principles adopted from Augustine's influential commentary on the Biblical account of creation in *Genesis* 1–3 (Fuller and Lipinska 2014, pp. 52–56): (1) a literal reading of humans having been created "in the image and likeness of God" (*imago dei*); (2) a stress on the qualified nature of God's forgiveness of Adam's "original sin" (*peccatum originale*); (3) the placement of equal emphasis on the perfection of divine creation as a whole and the radical imperfection of its parts (*theodicy*); and (4) a conception of God's creation of the world through the Word "out of nothing" (*creatio ex nihilo*). In this section, I focus on the paramount importance of the first principle (*imago dei*), postponing a more explicit discussion of its relationship to the other three until the next section. I preface my discussion of these principles here with a brief methodological note on how Fuller himself intends his "literal" reading of Scripture to be understood. For Fuller, a literal understanding of the Bible does not imply that we ought to read Biblical texts as historical documents; rather, it means treating them akin to a theatrical script (Fuller and Lipinska 2014, p. 53). That is, whereas a "literal reading" in the former sense would be governed by the ideal of *recovering* the Word of God (by finding out what the author meant back then, as determined by ecclesiastical authorities), the goal of Fuller's "literal readings" is to *re-enact* the Word of God (by figuring out for oneself what the author would mean now, within a specific contemporaneous context). Fuller has argued that these two different ways of reading Scripture, which nowadays are commonly associated with Catholic versus Protestant sensibilities, are prefigured by "Petrine" vs. "Pauline" approaches to apostolic ministry (Fuller 2008, Chapter 7; see Hirsch and Catchim 2012, Chapter 6).

In the Augustinian rendering of the Biblical account of creation, as recounted in *Genesis*, God brings the totality of the material world, directed to his purposes, into existence "out of nothing". It follows from this doctrine that everything that was created by God points to Him as its ultimate source of being, exhibiting its radical

[8] There is a growing number of articles, book chapters, and anthologies that engage Christian perspectives on transhumanism and transcendence; see, e.g., Brooke (2005), Cole-Turner (2011), Hansell and Grassie (2011), Delio (2014), Mercer and Trothen (2014), Malapi-Nelson (2017). Comparatively few authors have considered the affinities between transhumanism and the extended mind thesis; but see, e.g., Clark (2007), reprinted in More and Vita-More (2013). To my knowledge, no one has yet examined the theological moorings of the extended mind thesis, from a transhumanist angle.

contingency upon the exercise of His will. However, a central further element of Christian belief is the idea that human beings bear a special resemblance to God, by having been created in His image and likeness. Because of this similarity, humans enjoy a privileged standing within the cosmos that elevates them above all other created beings. Within a broadly Augustinian approach to Christianity, the communion between human and divine being is commonly expressed in the language of having shared access to the *logos*, in the sense that God creates through the Word, and we are uniquely positioned to understand His creation through the Word (Fuller 2015). The general terms of this mutual exchange warrant an attitude of confidence in human nature and a fundamentally optimistic assessment of our epistemic powers. It provides a metaphysical guarantee for the notion that the natural world – the *cosmos* – is in principle intelligible for creatures with the cognitive powers we have been granted – at least in a state of God's grace, even though the full exercise of these powers may be unavailable to humanity in its fallen state, as we shall discuss later. The project of articulating the precise nature of the attributes in terms of which human and divine natures can be said to be similar, or perhaps literally overlap, has had a profound impact on Western philosophy (Craig 1987).

Theologically speaking, the most dramatic expression of the communion between God and humanity is the Christian doctrine of *Incarnation*.[9] Within Chalcedonian Christology, human and divine nature are considered to form a "hypostatic union" in Christ – without confusion or separation ("two-natures, one person") – as the second Person (understood as individual substance or "supposit" of a rational nature) of the Trinity. This understanding raises a question which began to be intensely debated soon after the promulgation of the Chalcedonian "two-natures, one person" doctrine by the early Christian Church (Adams 1999): if Christ's *human* nature involves having both a human body and a human soul, then what exact form did the Incarnate God assume as a flesh-and-blood human being, i.e., prior to Jesus' death and resurrection? It appears that God could have taken up human form in four different conditions which correspond to different stages of human salvation – *pre-lapsarian* (Adam and Eve), *post-lapsarian* but before grace (Cain), *post-lapsarian* but with the help of grace (us), or as *glorified* bodies (and thus immutable and impassible, like after the resurrection). If each of these conditions is accidental and thus compatible with human nature, Christ could have assumed human form in any one of them, and be "fully" human *ante mortem*. Neither Scriptural exegesis nor conciliar pronouncements offered a principled answer to illuminate the speculative topic of Christ's human nature, so it became a matter of spirited theological debate.

As McCord Adams shows in her magisterial treatment, the range of variation among Chalcedonian portraits of Christ's human nature (ranging from Anselm to Luther) were driven by two principal considerations: to assess the purposes and proprieties of the Incarnation, and, based on this estimate, to work out an adequate "job description" for Christ's salvific work on Earth (Adams 1999, p. 9). The key

[9] For an insightful discussion of the Incarnation which aptly utilizes an "extended mind" metaphysics, see Marmodoro (2011).

strokes of each portrait are thus rendered in the attempt to balance out a variety of theological and philosophical motives (Adams 1999, p. 95). In response to inherited traditions that militate against the very idea that God would enter into a hypostatic union with a created nature, deeming it blasphemous and metaphysically incongruent, philosophers still felt the need to justify the Incarnation. This creates the "top-down" pressure to conceive of Christ's human nature as deiform as possible, hence to endow it with maximal perfection. This demand figures in theological appeals to *communicatio idiomatum*, the presumed communication of attributes between human and divine natures indwelling in the person of Christ, which was seen as ruling out certain attributes (e.g., the ability to sin) as incompatible with Jesus' divine nature. At the same time, the portrayal of Jesus' assumed nature was driven "bottom-up" by soteriological considerations. Specifically, the doctrine that Christ would assume something from each of the four conditions of human nature splits into two competing demands (ibid.). On the one hand, that Jesus' human nature be glorified from the moment of its creation, befitting his unique role as the divine savior of mankind. On the other hand, that he share enough mental and physical defects with fallen humanity (e.g., suffering and struggle) to furnish humanity with an inspiring role model of virtues that it ought to cultivate along the path to salvation.

Whichever way those two principal constraints were balanced, the triune conception of God as capable of taking up and entering distinctively human form certainly encourages the thought of a certain metaphysical commensurability between human and divine nature. Indeed, the Christian belief in the restorative power of the Incarnation is occasionally cast by saying that God became like us so that we might be made like God. We need to look no further than §460 of the *Catechism of the Catholic Church* (Catholic Church 2012) for textual evidence:

> "The Word became flesh to make us "partakers of the divine nature" (2 Pet. 1:4): "For this is why the Word became man, and the Son of God became the Son of man: so that man, by entering into communion with the Word and thus receiving divine sonship, might become a son of God" (St. Iranaeus, *adv. haeres* 3, 19, 1: PG 7/1, 939). "For the Son of God became man so that we might become God" (St. Athanasius, *De inc*., 54, 3: PG 25, 192B). "The only-begotten Son of God, wanting to make us sharers in his divinity, assumed our nature, so that he, made man, might make men gods" (St. Thomas Aquinas, *Opusc*., 57:1-4)" (CCC, §460)

In this "great exchange", as it was referred to by early Christian thinkers, God lowers Himself to man so as to redeem human nature. Should we understand this as implying that as humans, we are destined to become deified (*theosis*)? In the mainstream Christian churches of the Latin West, deification is emphatically understood in a modest fashion, as our human *participation in*, rather than our *possession of* a divine nature (Olson and Meconi 2016). That is, human beings are not meant to become equal to God, but rather have been adopted, through the Word of God,, to live, as "God's people," in His image and likeness, the mark of which is indwelling in the human soul. Going beyond this modest interpretation, Fuller has advocated a more ambitious reading of *imago dei* that takes as its starting point the "Transfiguration" (*theosis*) of Jesus, recounted in the New Testament as a miraculous episode during which Jesus caught a glimpse of his divine nature (Matthew

17:1–9; Mark 9:2–8; Luke 9:28–36). Fuller's strong reading of *theosis* envisions a more deeply transformative process, one in which ultimate goal for humans would be to literally achieve God-like status or even union with God – a theme that has been developed far more prominently within Eastern Orthodox Christian traditions (Burdett 2011, Fuller 2017). In those traditions, *theosis* is widely understood as a process of spiritual development that begins with the purification of mind and body (*catharsis*) to reach illumination with the vision of God (*theoria*), with the final goal of attaining sainthood and unification with God (*theosis*).

To be sure, it is true for mainstream Christian theologies of both Eastern and Western Churches that there remains an ontological separation between God and human beings, in particular insofar as Jesus Christ is seen as the Incarnation of a preexisting God that took on human form. Still, as the great sociologist of religion Max Weber observed, the distinctly Christian doctrine of Incarnation uniquely blurred the dividing line between human and divine, and effectively secured the Christian faithful a place at the divine table – whether this was understood modestly, as participating more fully and closely in the life of God, or ambitiously, as literally enabling men to become God-like (Weber 1963; see also Harari 2017). The Biblical Adam, having been created in the image and likeness of God, had already manifested our divine likeness in its original, unadulterated form, but had to forfeit many of his God-like characteristics (both physical and mental) as a result of the Fall. However, once Jesus (identified by Paul as the "last Adam") arrived on the scene to undo the debilitating effects of Adam's ill-fated transgression, the prospect of recovering our lost "Adamic" perfections was at least in principle made available to all Christian followers, and became an integral part of the Christian salvation project (Noble 1997). Its most explicit scriptural expression can be found in the Millenarian Book of Revelation. It prophesized a "happy ending" to the Biblical account of anthropogenesis which promised the faithful a return to the way things once were in the Garden of Eden, with human mental and physical powers fully restored to their God-like perfections, and with human dominion over nature fully reestablished.

From this soteriological angle, Fuller conceives of modern science and technology as two complementary modes of human deification (*theosis*) – i.e., as ways of "getting into the mind of God" (via science, in particular mathematical physics) and of "playing God" (via engineering, in particular the development of nano-, bio-, information-, and cognitive technologies). More specifically, Fuller links the transformative potential of science and technology to two of the most enduring Christian heresies (Fuller and Lipinska, Chapter 2): *Arianism* and *Pelagianism*, each named after their originators, Pelagius (a Celtic lawyer), and Arius (a Libyan bishop). Common to both heresies is a kind of "do-it-yourself" conception of salvation that accords humanity sufficient redemptive powers to raise itself from its fallen condition, to get back into God's good graces (although God may well retain the last word on whether our efforts have been successful). Grounding this rather sanguine assessment of human potential is Fuller's second aforementioned theological principle – a specific gloss on the doctrine of original sin. Fuller interprets it as saying that with his death, Jesus effectively nullified the lingering effects of Adam's original sin, fulfilling his divine mandate by putting us back on track to realize our full potential

if only we follow in his lead. Adopting a Unitarian Christian perspective, Fuller sees Jesus as a divinely inspired (and in that sense exceptional) yet still human "role model" – a moral authority with the power to inspire others to lead similar lives, in their own times, without any further need for clerical approval. Against Trinitarian Christians who accuse Unitarians for their denial of the divine nature of Jesus, Fuller would retort that what Unitarians deny is merely the *uniqueness* of Jesus' divinity (Fuller 2011, p. 98).

If science and technology are prefigured by Arianism and Pelagianism, respectively, as two different strategies for elevating humanity to divine heights, how do they differ from one another? The Arian heresy is derived from Arius' apparent denial that Jesus was of the same Nature ("con-substantial") as God, against which he asserted the nontrinitarian belief that Christ was begotten by God and is thus subordinate to Him. This difference in ontological rank means that Jesus didn't inherently enjoy a special metaphysical relationship to God that other humans – *qua* having being created in *imago dei* – necessarily lack; it's just that Jesus discovered how each of us might be able to enter into such a relationship, thereby fulfilling our quest to achieve reunification with God. From a broadly Arian perspective, *theosis* thus refers to a process of spiritual enlightenment through which humanity quite literally comes to realize its divine nature, something it only imperfectly recognizes in its fallen state. Overcoming this deficit thus requires the pursuit of a variety of "idealistic" projects for disciplining the mind, by freeing it from its imprisonment in a body which hinders the full exercise of our mental powers. The archetypical expression of this viewpoint is Plato's metaphysical dualism.

In Fuller's far-ranging historic treatment, the quintessentially "Arian" strategy for achieving salvation is shown to have exercised a major influence on subsequent philosophical developments. Those include the medieval development of Platonistic Christianity, in which Plato's eternal Forms are equated with the thoughts of God, to the effect that human and divine minds can be said to overlap in their grasp of these eternal truths. As Minister General of the Franciscan order, Bonaventure took his mystical vision of our *Journey of the Mind to God* as a metaphysical template for the idea of a "course of study" (aka curriculum) at the nascent medieval universities. In the hands of early modern philosophers, the human potential for being able to access the mind of God mind gave rise to the rationalist conceptions of *a priori* knowledge and *innate ideas*. Newton conceived of his epistemological ideal, the infamous "view from nowhere" (Nagel 1986), as a human simulation of the standpoint of God, whose omniscience Newton explained as a function of God's being equidistant from all points in space and time. Fuller has argued that the detached, objective perspective which is bound up with taking the stance of Newtonian physical science (*theoria*) is a secular descendant of the "Transfiguration" of Jesus – i.e., the synthetic ability to transcend human partiality by viewing the entirety of Creation ("Nature") from a God-like perspective (Fuller 2011, 2015, 2019). It is well-documented that Newton's heterodox Arian views nearly derailed his academic career (Westfall 2007).

In contrast, the Pelagian heresy proscribes a more "materialistic" stance to promote the divine aspirations of humanity, which originates in the dissident Celtic

monk Pelagius' rejection of the Augustinian emphasis on divine grace as a vehicle of human redemption. Augustine had insisted that moral perfectibility is beyond human control in its fallen state. Whatever moral or spiritual progress can thus be made during the span of our earthly existence can be achieved only with the help of God's grace, which He bestows upon us as His gift. Against this doctrine, which he perceived as a troublesome source of "moral laxity," Pelagius denied that human nature was inherently corrupted by Adam's sin. Humans are born with the capacity to perfect themselves, i.e., to follow God's commands, by the exercise of their free will, and thus without having to rely on the aid of divine grace; although God will no doubt assist those who do the right thing. Stressing the Pelagian emphasis of the independent ability of humans to act morally, and to strive towards salvation, Fuller associates the Pelagian heresy with various "materialist" projects of transforming and reshaping the material conditions of human life to extend our mental powers, and more generally to further human intents and purposes. The archetypical model for this "materialistic" approach to salvation, for Fuller, unfolds as the history of human technology.

Here, Fuller points to the work of historian David Noble (1997), who showed how technology in the West became invested with spiritual significance as a means of restoring humanity to the state of its original, "Adamic" perfection, with the possibility of renewing that perfect state on Earth. The Carolingian philosopher John Scotus Eriugena, for example, coined the term "mechanical arts" – a precursor of the term "useful arts" and later "technology" – and dignified their pursuit as a means for recovering the *corporeal* respects of our divine likeness which were lost as a result of the Fall. Often hailed as a technological visionary who was deeply moored in the Joachimite millenarian milieu of his time, Roger Bacon considered the advancement of technology as contributing doubly to the human quest for salvation: first, as a tool for recovering the lost Adamic knowledge of Nature that once was part of our natural endowment; second, in anticipation of the second coming of Christ, as preparing the Christian faithful to achieve victory in their impending battle against the Antichrist. In the seventeenth century, during which the King James Bible vividly made such apocalyptic prophecies fit for public consumption, the same millenarian zeal galvanized the Reformed Christians who spearheaded the Scientific Revolution, most notably Francis Bacon (Webster 1975). Within their decidedly pragmatist outlook, the portrayal of Adam as an all-knowing and all-capable artisan served as a Biblical source of great inspiration (Noble 1997, Chapters 1–4).

A pivotal moment in the historical development of distinctively "modern" conceptions of progress, including Enlightenment conceptions of science and technology as redemptive forces, is the medieval debate over the nature of divine predication, i.e., the legitimacy of human ways of talking about God (Fuller 2011, Chapter 2). For example, when we assert that God – considered as a supremely perfect being – is "all-powerful", "all-good" or "all-knowing", are we using the terms "power", "good" and "knowledge" in the same sense in which we apply them to humans? In the scholastic literature, the dispute between Thomists and Scotists over the nature of divine predication comes down to this. Following Aquinas, Thomists hold that

when we apply these predicates to God, we only apply them *analogically* to the ways in which we apply them to His creatures. Ontologically speaking, this linguistic division multiplies the realms of being and knowledge which separate us from God. In contrast, following Aquinas' great antagonist, Duns Scotus, Scotists argue that we apply these attributes *univocally* to all beings, i.e., we apply them in the very same sense to both God and His creatures. Ontologically, this means that the difference between us and God is one in degree rather than in kind. For Scotus, the concept of "infinite being" – understood positively, as an intrinsic degree of perfection, and not negatively, as a limitation – is the simplest possible concept which human beings can acquire that also univocally applies to God (Williams 2016). As Fuller points out, taking a Scotist line on divine predication not only allows for direct comparisons between human beings and the deity along a single continuum, but actively encourages the idea that the ontological distance between humans and God is in principle reducible by making human attributes more God-like, i.e., by *improving* them indefinitely. Historically, modern sensibilities have generally followed the Scotists in this regard. During the Enlightenment, "humanity" became gradually viewed as a species-wide project of self-transformation for which the dynamic ideal of "perfectibility" (Passmore 1970) – i.e., the capacity to be improved to an unlimited degree – provides a moral imperative to restore our fallen, imperfectly embodied selves in the "image and likeness" of God (Funkenstein 1986; Harrison 2007). Another secular reflex of the Scotist position is what philosophers of science call "convergent realism" – the conviction that an ultimately completed edifice of science would provide a true and comprehensive representation of reality (Fuller 2011).

Even though the Scotist doctrine of univocity is a semantic thesis, it suggest a particular image of the post-lapsarian predicament in which the human species finds itself (Fuller 2015, p. 98). This image, also Scotist in its provenance, describes the punitive damages that God inflicted on our godlike capacities as a result of the fall as a loss of *unity* (or "integrity") among the virtues which in God's divine nature are harmoniously integrated. In other words, whereas God is the one being in whom all virtues are concentrated perfectly, in one divine person, it is plain for all to see that those virtues are distributed rather imperfectly among many human individuals. That is, the most knowledgeable among us are not necessarily the most powerful, the most powerful are certainly not always morally upstanding, and even the most knowledgeable remain capable of committing great evil. Thus, the Scotist rendering of the Fall depicts the lingering effects of God's punishment as "a dispersion of the self rather than a collective demotion of the species" (Fuller 2015, p. 98), i.e., the stunted and unequal development of virtues we observe among people, without a clear sense of how they might be put back together again to the collective benefit of humankind. In *Genesis*, the resultant state of dispersal is captured in a linguistic simile – the "Tower of Babel" narrative, which tells the story of an alienated humanity divided against itself. For Fuller, the Scotist imagery prefigures the Enlightenment ideal of "humanity" as an ongoing remedial project which calls for the re-integration of our woefully dispersed virtues, both within and among individuals. Exactly how this collective project ought to be organized, with the highest chance of success, defines and at the same time divides the modern political spectrum (Fuller 2011,

Chapter 2; Passmore 1970). Consistent with his Unitarian Christian stance, Fuller holds that not even Jesus managed to fully integrate all of the divine virtues in his existence as a person, although the level of unification he achieved was sufficiently exceptional to inspire in his followers a desire to follow his lead (Fuller 2012, p. 98).

To conclude this section, I quote Fuller's summary of how science and technology constitute complementary pathways towards human salvation:

> "In a nutshell, Pelagians imagined a "heaven on Earth," whereas Arians imagined an "Earth in heaven." These alternative visions have resonated with the process of modernization. Thus, the technological transformation of the life-world to maximize human convenience is a Pelagian project, just as the scientific aspiration for a maximally comprehensive theory of reality (aka "entering the mind of God" or the "view from nowhere") is an Arian project. Whereas the Pelagian aims to reduce the time it takes to realize the human will, the Arian aims to expand indefinitely the scope of humanity's intellectual horizons. In the modern era, the two movements worked in tandem. The seventeenth-century scientific revolution in Europe marked the triumph of the Arian vision, on the basis of which the eighteenth-century Industrial Revolution began to make the Pelagian vision a reality" (Fuller 2017, p. 384).

In Fuller's vision, the difference between a Pelagian and an Arian conception of salvation is cashed out as two modes of human transcendence, understood as self-enabled albeit divinely inspired attempts through which human beings seek to realize their god-like potential:

> Thus, instead of trying to *humanize nature*, *à la* Pelagius, the Arians wanted to *denaturalize humanity*. The former is epitomized by 'biomimetic' projects that enable the extension of human powers (including longevity) by incorporating elements of the rest of nature (Benyus 1997). The latter is captured by what transhumanists have called 'morphological freedom' – that is, the prospect of the human essence migrating across material forms, including from carbon to silicon incarnations and possibly pervading the entire cosmos, *à la* Kurzweil (2005)." (Fuller and Lipinska 2014, p. 49).

Having reviewed the outline of Fuller's transhumanist project, I now return to my thesis that we can profitably conceptualize important tenets of Clark's EMT as a chapter in the larger history of transhumanist thought. Here is how I intend to argue for this claim. For our purposes, I shall leave untouched the historical arc of Fuller's narrative in support of his heterodox Christian development of transhumanism, and concentrate on showing how the EMT fits into this picture. Specifically, I argue that central themes of Clark's thesis can be seen as materialistically reworked descendants of the *imago dei* doctrine, as well as related theological principles as mobilized by Fuller. Since the themes to which I shall refer are closely connected in Clark's own presentations of the EMT, but also within Fuller's transhumanism, they will need to be disentangled for analytic purposes over the course of the next two sections. While fixating on the overarching impact of the *imago dei* doctrine, I organize my discussion around three historically influential ways in which this doctrine was philosophically articulated in the modern period (cf. Craig 1987): the early modern "ideal insight" and "ideal agency" models (Sect. 13.3), and the romanticist "practice" model (Sect. 13.4).

As a final caveat before proceeding, I ought to stress that for my historical reconstruction to succeed, and have genuine interpretive value, the reader need not

subscribe to the substance of either Clark's EMT or Fuller's transhumanism; nor should we assume that these two thinkers would necessarily assent to each other's viewpoints. That is, I clearly do not mean to assert that Clark is actively pursuing a clandestine theological agenda in support of Fuller's transhumanism; nor do I wish to imply that Fuller himself would find the EMT to be a congenial ally to further his transhumanist ambitions. Instead, my argumentative strategy is "archeological" in the sense that I seek to unearth and trace out the subterranean impact of recognizably transhumanist motives and aspirations that I find to be present in Clark's EMT, but whose theological provenance has mostly gone unrecognized by its main proponents.

13.3 Extended Minds Aspiring to the "Ideal Insight" and "Ideal Agency" Models of Imago Dei

Discussions of the "mind-body problem" in early modern philosophy can always be seen as proxies for competing philosophical assessments of our metaphysical standing, as human beings, on a continuum between Heaven and Earth. Within that "Great Chain of Being" (Lovejoy 1936), humankind traditionally occupies a special position that straddles the realms of purely spiritual beings and the natural world. In particular, what is taken to set us apart from other living creatures, within this Chain, are the distinctive powers of the human mind, such as our creativity, the intellect, or the will; in contrast, our animal bodies betoken our connection to the lower ranks of this ontological hierarchy. Concerning our standing in this Great Chain, Fuller speaks of a "bipolar disorder" that is running through Western conceptions of the human (2012, Chapter 2): are we by nature spirits, and thus only contingently tied to the possession of animal bodies (*dualism*); or are we fundamentally animal bodies, albeit with special kinds of minds, courtesy of our overdeveloped neocortex (*materialism*)? The perceived standing along this great divide determines whether humanity's ontological kinship – and thus primary sense of affiliation – lies with a transcendent *Deity* or immanent *Nature*. In their overall assessments of what it means to be human, transhumanists veer towards the former, whereas posthumanists side with the latter.

Within the Christian tradition, the doctrine of *imago dei* has been invoked as a divine mandate for human exceptionalism, thus underwriting a characteristic attitude of confidence in the human intellect, its powers to know, and our ability to establish dominion over nature. But as a theological motto, the idea that man was made in the image of God is rather indeterminate, and does not amount to much of a distinctive philosophical thesis. Early modern philosophers thus labored hard to make its content more precise, by articulating specific dimensions in which human and divine minds might be literally said to be alike or indeed overlapping – "perhaps in the same sense as a human and a chimpanzee genome 'substantially overlap'" (Fuller 2015, p. 75). Closely following Craig (1987), I briefly recount the two main

models of how the notion of *imago dei* was taken up by early modern philosophers.[10] In my discussion, I connect these two models to the other theological principles, mentioned earlier, which Fuller has mobilized for his transhumanist project. I then argue, from a Fullerian vantage point, that the human drive towards cognitive extension is born out of necessity as our "post-lapsarian" minds – encumbered by the "design flaws" of our evolved biological brains – are struggling to live up to the lofty standards set by the two early modern models of *imago dei*.

To begin with, let me restate the philosophical conundrum which early modern philosophers had to face in their attempt at fleshing out the exact content of the *imago dei* doctrine. Theologically speaking, if we accept the Scotist univocity of divine predication, it becomes particularly puzzling how a literal understanding of our likeness with God can be reconciled with the doctrine of divine perfection. With respect to any mental power or cognitive virtue that we might deem human beings as capable of having, it seems clear that current human performance levels fall dramatically short of the divine benchmark of perfection. As created, and thus by nature finite beings, how could we possess, or even coherently aspire to achieve mental powers that would literally be God-like? As Craig points out (1987, Chapter 1), early modern philosophers sought to bridge the gap between human and divine nature in terms of two main models: the dominant "ideal insight" model of human perfection, and the less prominent yet still influential "ideal agency" model of perfection. Let me briefly characterize each model in turn.

The "ideal insight" model turns on those aspects of human *knowledge* (or *cognition*) which can coherently be extrapolated to perfection. What might those be? In his *Dialogue Concerning the Two Chief World Systems*, Galileo (1632/2001) drew a distinction between the *extensive* and *intensive* aspects of human cognition which proved to be extremely influential in this regard. From an *extensive* or purely quantitative standpoint, an omniscient God knows everything there is to know, which would presumably include an actual infinite number of truths. Since the capacity of the human intellect, in virtue of having been created, is necessarily finite, having perfectly extensive knowledge must forever remain beyond our ken. However, Galileo argues that things look more promising if we consider the modes in which humans acquire and possess knowledge, i.e., knowledge from an *intensive* or qualitative standpoint.

If having knowledge is a matter of reliably acquiring beliefs that are true, an ideal knower would be one who is infallible, and has utmost *confidence* in the beliefs she holds to be true (aka *certainty*). A related qualitative benchmark of epistemic

[10] Craig (1987, 15f) mentions three factors that would have contributed to the Early Modern preoccupation with *imago dei*: first, the rise of Protestantism, with its emphasis on the closeness between God and individual human beings, such as the call for the individual to discern, without reliance on ecclesiastical mediation, God as the author of the book of Scripture and the book of Nature (cf. Harrison 1998); second, a kind of unconscious defense mechanism against the discoveries of the new sciences, specifically their threat to displace humanity from the center of cosmic significance; third, a reverence for Christian ideas that could be philosophically sustained without deference to traditional authorities.

perfection would be the *manner* in which knowledge is attained. For example, if we value the speed and ease of knowledge acquisition, an ideal knower would be one who apprehends all things instantaneously and effortlessly. Alternatively, if perfection is a matter of exercising voluntary control over one's thought processes, an ideal knower would be one who creates knowledge freely, without being subject to any external constraints. Consistent with the doctrine of divine perfection, it should also be noted that both "ideal insight" ideals have an "intrinsic maximum", i.e., a logically coherent level of maximal perfection that is proper to the divine mind; albeit one that human minds can aspire to emulate, notwithstanding their finite nature. In addition, since God is also supremely good, seeking to acquire ideal knowledge of this kind would naturally be seen as a display of moral excellence – because its attainment constitutes a good thing in itself, regardless of whatever ends to which it may be applied. Moreover, it puts human beings under a moral obligation to pursue knowledge endeavors which befit their God-like capacities, and we can fulfill these obligations by continually seeking to improve our epistemic faculties with respect to the relevant dimensions. In this sense, the contemplative "ideal insight" model of human epistemic perfection can be seen as a continuation of the Arian conception of *theoria* as *theosis*.

The "ideal insight" model can be articulated according to the demands of an empiricist or a rationalist epistemology (Craig 1987, pp. 27–44). For empiricists, what primarily qualifies as an "ideal insight" is the immediacy and transparency by which we know what's going on in our own minds. For example, Newton compared our conscious introspective awareness of mental images to God's own awareness of the universe *sub specie aeternitate*, considering space as the divine "sensorium" through which God has perfect knowledge of the physical world. Similarly, Berkeley thought that human self-knowledge provides such an excellent model of epistemic perfection that he proceeded to analyze physical objects as ideas in one's mind (*esse est percipi*). This not only brought human knowledge of the physical world into the ambit of infallibility and certainty, but also had the effect of assimilating human acts of *perceiving* the physical world to divine acts of *creating* it – on both occasions, creation would be likened to a voluntary act of putting ideas into one's minds.

For rationalists, the "insight ideal" model is best exemplified when we grasp the necessity of certain truths, such as the eternal truths of logic and mathematics. As noted earlier, this model goes back to quasi-mathematical renderings of the Platonic doctrine of participation, and its privileging of *intuitive* (immediate) over merely *discursive* (mediated) modes of knowledge. Historically, the ideal of modern science as an axiomatic, deductively closed system whose inferentially interconnected truths can in principle be apprehended with the same certainty and immediacy grew out of this variant of the "ideal insight" model of perfection. The advances of modern natural science, in particular fundamental physics, towards conforming to this ideal would thus be evidence for both the God-like powers of the human intellect and the rational intelligibility of the world (Fuller 2015).

The "ideal agency" model of human perfection, which was far less prevalent among early modern philosophers, is concerned not with knowledge but the powers

of the will.[11] For Descartes, the signature trait of humanity which underwrites our divine potential is the infinite freedom of the human will, defined by Descartes as the capacity to choose in ways entirely unconstrained by forces external to the mind. As Descartes argues in the *Fourth Meditation*, not even God could be any freer in this regard, although His will is no doubt incomparably greater in efficacy and scope. With his voluntarist conception of free will, but also by endorsing a metaphysics of modality in which possibility is viewed as whatever is logically conceivable, Descartes belongs to an intellectual tradition that can be traced back to Scotus, and which Fuller regards as a precursor to contemporary transhumanism (Fuller 2011, Chapter 2.3; see also Shiffman 2015). In opposition to Aquinas, Scotus had similarly defended a voluntarist conception of free will as literally God-like in its radically unconstrained freedom to create, which would secure a metaphysically exceptional status for human beings vis-à-vis animals.

A sophisticated and powerful articulation of the "agency ideal" of *imago dei* can be found in the work of Leibniz. For Leibniz, the perfect agent would be one who always chooses the right goals, while being entirely unconstrained in these choices, and who always succeeds in achieving those goals. To be sure, this purely functional definition doesn't tell us what kind of calculus a perfect agent might employ to arrive at her decisions, which was the subject of much philosophical debate. From the standpoint of *imago dei*, though, it raises the question of why human agency visibly falls short of this divine ideal. Because even if we grant that our actions are chosen freely, we plainly do not always set ourselves the right goals, nor do we always make the right plans to bring them about. Leibniz explained the shortcomings of human agency as a purely *epistemic* limitation – an exercise of poor judgment on our part, rendered under the influence of our fallen condition. This may seem like a psychologically implausible idealization; but it allowed Leibniz to maintain, appearances notwithstanding, that we really do act like the deity at least insofar as we always pursue goals that *seem* right to us, and always choose a course of action which *appears* most appropriate to us, although we don't always (or perhaps only rarely) get things right (seen from a divine standpoint).

For Leibniz, this essentially epistemic solution assumed great metaphysical significance in the context of addressing the problem of evil. This takes us to the third theological principle undergirding Fuller's transhumanism: the placement of equal emphasis on the perfection of divine creation as a whole and the radical imperfection of its parts (Fuller and Lipinska 2014, p. 54). Traditionally, this principle has been invoked in the service of *theodicy* ("divine justice"), a term coined by Leibniz to denote philosophical attempts of explaining why an all-mighty God who is also all-good and all-knowing could permit the manifestation of evil, such as human suffering or the preventable occurrence of horrific natural catastrophes. Adopting a Leibnizian standpoint, Fuller takes seriously the idea that even God can only imperfectly cope with the intrinsic limitations of the material medium in which His

[11] In the next section, we shall turn to the "practice" ideal of *imago dei*, a historically subsequent development which nevertheless has important features in common with the earlier "agency" ideal (Craig 1987, Chap. 5).

creation must take place. He adopts from Herbert Simon's (1981) account of "bounded rationality" the notion of a "constrained maximizer" to describe the dilemma in which a Leibnizian creator of the world would find himself. Aiming for the best possible overall outcome, but constrained by the imperfections of matter, the problem of creation requires God to make calculated trade-offs that inevitably lead to outcomes which, judged from a local perspective, appear less than optimal. Those outcomes are what we regard as evil from a human standpoint, but they are motivated, in fact dictated (and thus justified, or so the theodicy goes) by God's benevolent goal of executing the best possible overall plan. Thus, according to Fuller's "bounded rationality" interpretation of the "ideal agency" model, the Leibnizian will can be characterized as a cognitive faculty – equally present in God and in humans – that is capable of performing the calculations that are necessary to engage in such boundedly rational decision-making.

Appealing to a similar logic, Fuller and Lipinska (2014, p. 54) go on to argue that God's reason for forbidding Adam and Eve to eat from the Tree of Knowledge was his wish to protect humanity from the "dirty work" that inevitably comes with creation – i.e., possibilities that have to be foregone in order to enable things to be as they are. Once God had punished humans for Adam's sinful deed, the only path for humankind to recover the divine plan, to which the forbidden tree held access, would be through our own efforts, however imperfect those may be. For Fuller, scientific inquiry represents the best shot we have at "reverse-engineering" God's plan, by aiming to understanding the universal laws governing His creation in the most economical terms. Fuller thus takes the lingering epistemic effects of the fall as a rationale for endorsing the theory of intelligent design (Fuller 2008).

The fourth theological principle which Fuller has mobilized as a foundation for his transhumanism is the idea, common to the Abrahamic religions, that God creates the world in a free and unconstrained act "out of nothing" (*creatio ex nihilo*). Fuller comprehends this unique display of divine agency as a paragon of efficiency which human beings can strive to emulate. Of course, unlike God, human beings cannot literally create something out of nothing, but they can approximate the divine ideal through *economizing*, i.e., by using smaller and smaller amounts of resources and labor to accomplish more and better outcomes. As utility-maximizing agents, we can achieve efficiency gains in the work we perform in two distinct but complementary ways (Fuller and Lipinska 2014, p. 55). On the input side, we can increase efficiency by diminishing the marginal cost of units produced, i.e., by reducing the total cost of labor and/or resources that are necessary to increment the quantity already produced by one. Conceived abstractly, the principle of reducing marginal costs applies not only to the production of material but equally to ideational goods. An example would be an appeal to "Occam's Razor" in favor of ontological parsimony, which is frequently employed in the context of scientific reasoning. As Fuller has argued, if the natural world bears signs of the craftsmanship of an intelligent designer, the principle of "Occam's Razor" can be justified on the grounds that God, as a maximally efficient creator, would only choose to bring into existence the minimal number of ontological building blocks from which everything else can be derived (Fuller 2008).

On the output side, efficiency can be gained by increasing the rate of return, which means that greater productivity can be achieved with less effort, time, or energy. With recourse to this principle, Fuller has promoted the human value of "ephemeralization" – a phrase coined by the Unitarian inventor, futurist, and systems theorist Buckminster "Bucky" Fuller (1938/1971) to describe the notable trend in human history towards "progressively doing more with less". Bucky's point was not only that can we increasingly do more with less, but that the rate of doing-more-with-less-ness is increasing. In his book, he gives many examples of the accelerating rate of technological progress. For example, it took Magellan two years to sail around the planet in a wooden sailing ship in 1520; 350 years later it took a steel steamship two months to do the same; 75 years later a plane, made of metal alloys, took two weeks to fly around the planet; 35 years later – and thus around the time his book was written – it took a space capsule, made of exotic metals, a mere hour to circle our planet. The materials we use to construct such transportation devices have gotten progressively lighter, stronger, and more versatile. Today, the reliability of "Moore's Law" (Moore 1965), which roughly predicts that our available computing power doubles approximately every two years, is frequently cited in support of a continuation of this universal trend. Since the process of ephemeralization does not seem to have an inherent upper bound, we can extrapolate a God-like ability to do everything with nothing as its asymptotic limit.

This concludes my thumbnail sketch of the "ideal insight" and "ideal agency" models through which early modern philosophers sought to articulate the content of *imago dei*.[12] What I now intend to show is that inflected developments of these two models continue to thrive, with a decidedly materialistic twist, within Clark's EMT. Earlier, we said that within the Western tradition, the doctrine of *imago dei* has been widely appealed to as a justification of human exceptionalism. In this context, I noted earlier (Sect. 13.2) how Clark's image of the human as a "natural-born cyborg" preserves a distinctive commitment to the uniqueness of human vis-à-vis animal minds, but in a peculiar way which combines the dualist motif of transcendence with the materialist motif of incorporation. To prepare the ground for my argument that there is an intriguing intellectual continuity between early modern conceptions of *imago dei* and Clark's image of the human as a natural-born cyborg, we must first consider how Clark's EMT straddles the familiar epistemological divide between rationalist and empiricist approaches to human knowledge.

My access point are the opening passages in Chapter 8 of Clark's (2013) *Mindware*, in which Clark obliquely revisits the question of human exceptionalism by way of introducing his conception of "cognitive technology." In the discussion

[12] Fuller has fashioned a brand of social epistemology which aspires to blend these two models. He advocates a "constructivist" (more precisely: "realizationist"; cf. Fuller 2012, p. 272) and "agent-oriented" understanding of social epistemology in that it emphasizes knowledge as a form of *doing* "concerned with the *ethics* and the *engineering* of reality construction" (Fuller 2015, p. 15), as opposed to the "spectatorial" and "object-oriented" orientation that he associates with mainstream analytic epistemology. For an overview of Fuller's conception, see Remedios (2003), and Remedios and Dusek (2018, Chap. 3).

leading up to this chapter, Clark gave ample evidence of psychological mechanisms that are widely shared across species, such as parallel distributed processing in neural networks, sensorimotor coordination, and the dynamics of agent-environment couplings. At this point, the suggestion that there may be something special about human minds would seem to be particularly tenuous for proponents of situated and embodied cognition who – like Clark – stress the mechanistic continuity of our psychological equipment with that of animal minds. How, then, might we hold on to the notion that there may be a "special *kind* of mindfulness associated with the distinctive, top-level achievements of the human species", among which Clark characteristically cites "abstract thought, advance planning, hypothetical reason, slow deliberation" (Clark 2013, pp. 166–67). Historically, harping on precisely these features of human cognition would have been grist to the rationalist mill – a way of stressing the uniqueness of human reason. If we look back at the early history of artificial intelligence and cognitive science, it is thus hardly surprising that the promise of "physical symbol systems" (Newell and Simon 1976) to carry out such ratiocinative tasks commanded the attention of rationalistically minded philosophers (à la Jerry Fodor). But having spent the majority of his book distancing himself from the legacy of symbol-based computational understandings of cognition, Clark's empiricist gambit is to acknowledge the significance of these "top-level" achievements of the human mind, yet to explain them in a very different way. He sets out three options for doing so (see also Clark 2001).

The first option, which must count as a variant of rationalism in this context, is to posit a "deeply hybrid" cognitive architecture of the human brain. Here, a peripheral processing layer would be the locus of "quick and dirty," computationally cheap routines which are good at coping with the demands of sensorimotor processing; in contrast, only a core processing layer – factory-primed to engage in symbolic computation-– would be equipped to handle the demands of higher-level cognition. In more general terms, the strategy behind this view is to explain the cognitive *discontinuities* between humans and their closest kin in terms of underlying biological – specifically neural – *discontinuities*. The second option, a descendant of classical empiricist viewpoints, is to deny that there is any such stark psychological discontinuity between human and animal minds. Proponents of this perspective would insist that a few relatively minor tweaks to the same old neural architecture of our animal minds were sufficient to give rise to the full gamut of human rationality. In essence, the second option asserts a fundamental cognitive *continuity* between human and animal minds that is consistent with their deep biological *continuity*.

Rejecting the first two options, Clark presents his "tool-centered" account, which foregrounds the cumulative effects of technological scaffolding for the cognitive explosion in our lineage. This third option is a subtle blend between the first two insofar as it aims to reconcile a deep biological *continuity* with an equally profound cognitive *discontinuity* between humans and their closest kin. Unlike any other biological creature, says Clark, we surround ourselves with multiple layers of external and/or artificial cognitive aids ("wideware") in ways that continually re-shape and expand the space of human reason. Thanks to this increasingly potent co-evolutionary spiral, the enculturated and massively reengineered cognitive profiles of our

extended *minds* spectacularly exceed the abilities which would be endemic to our biological *brains*. From an explanatory standpoint, the third option is attractive, because it promises a parsimonious account of the distinctively human "top-level" achievements of human thought and reason, yet without positing any radical evolutionary leaps (Clark 2008, p. 57).

Clark's third option trades on a "strange inversion of reasoning" in the sense of Dennett (2009). Dennett himself borrows the phrase from R.B. Mackenzie, a nineteenth century critic of Darwinian evolution who mocked Darwin's suggestion that "to make a perfect and beautiful machine, it is not requisite to know how to make it" as a patently absurd denial of the role played by an intelligent designer. Having similarly stressed the transformative impact of "tools for thinking" (Dennett 1992, 1996), Dennett goes on in his paper to express his admiration for Alan Turing's notion of *computation* as another instance of a "strange inversion." Building on Dennett's own discussion, we can rephrase the central insight of Clark's EMT as follows: *In order to build a cognitive engine that is capable of satisfying the exacting demands of advanced human thought and reason, it is not necessary that the biological portion of such an engine must operate on the principles by which the (environmentally extended) engine abides.* Framing this motto in more traditional epistemological terms, I shall now argue that Clark's EMT, in particular its portrayal of human minds as "natural-born cyborgs" takes up central features of the "ideal insight" and "ideal agency" models of *imago dei,* albeit materialistically repurposed and considerably "downsized" to fit our extended human minds. Unpacking and substantiating this claim will take up the remainder of this section.

To add a concrete historical context, I find it instructive to compare Clark's emphasis on the precarious epistemic situation in which the situated human cognizer finds herself with the intellectual climate of seventeenth-century England, in particular the distinctly pessimistic view of the powers of the human mind in the aftermath of Adam's Fall – an assessment which motivated the rise of modern science (Harrison 2007). In this period, the encyclopedic and God-like intellectual capabilities that Adam would have originally enjoyed – having been created in the image and likeness of God – served as a standard of human perfection against which the severe limits of human knowledge in our time could be ascertained. For instance, Adam's unbounded knowledge would have meant that "he could view Essences in themselves, and read Forms with the comment of their respective Properties; he could see Consequents yet dormant in their principles, and effects yet unborn in the Womb of their causes" (South 1679, pp. 127–28; cited after Harrison 2007, p. 1). Joseph Glanvill, a founding member and leading propagandist of the Royal Society, recalls the great distance separating ourselves from the impeccable state of Adam's cognitive and also perceptual equipment: "We are not now like the creature we were made […] The senses, the Soul's windows, were without any spot or opacity […] Adam needed no spectacles. The acuteness of his natural optics showed him most of the celestial significance and bravery without a Galileo's tube […] His naked eyes could reach near as much as the upper world, as we with all the advantages of arts […] His knowledge was completely built […] While man knew no sin, he was ignorant of nothing else." (Glanvill 1931, pp. 3–5; 6; 8; cited after Noble 1997,

p. 61). For the collective imagination of seventeenth-century England, such effusive praises of Adam's pre-lapsarian perfections were hopeful reminders of an unfulfilled human potential that might one day be unlocked, specifically through the development and use of modern experimental science.

While not everybody shared Glanvill's sanguine belief that the new scientific method would allow for a complete recovery of Adamic perfection, it was nevertheless widely agreed that a precise understanding of the debilitating impact which the Fall had on the powers of the human mind (and body) would be needed to determine the rightful course for the advancement of knowledge. This common presupposition is noteworthy in several regards. From an ethical standpoint, referring to pre-lapsarian standards of Adamic perfection as the default state of human nature implicitly shifts the criterion for distinguishing interventions into our current state that are performed with *therapeutic* intent from those performed with the express intent of *enhancement* (Harrison 2007). That is, compared to the unblemished state of Adam's natural endowment, our fallen minds and bodies are "always already" disabled, and thus stand in dire need of *repair* and *restoration*. If we apply Glanvill's logic, then wearing glasses, peeking through Galileo's tube, or other forms of technologically extended cognition should count as *therapeutic* interventions to lift the fallen state of our imperfectly embodied minds back up to Adamic standards of perfection. Among early modern scientists, this essentially restorative project left the door open for "rationalist" aspirations to attain an even higher degree of similarity or identification with the mind of God, which would go beyond the recovery of Adamic God-likeness. As discussed in the previous section, this "supra-Adamic" ideal of perfection became associated with the "Arian" strategy of coming to know the mind of God through scientifically deciphering the divine design behind nature – an attitude that was perhaps most evidently motivating the scientific work of Newton and Boyle (Funkenstein 1986; Noble 1997).

From a philosophical standpoint, the religious background of early modern discussions of the Fall set the epistemological agenda for the next two centuries, in the following sense: negatively put, since epistemological error was equated with sin, the prevention of error requires that we correctly identify the causes of human ignorance and defect; positively put, a sober assessment of the nature and amount of degradation that God inflicted on humanity would be necessary to determine what kinds of things can in principle be known to fallen creatures, and by which methods. Consequently, as Harrison (2007) shows in great detail, different estimates of the severity of the Fall, and competing diagnoses of exactly which parts of human nature it corrupted, can be correlated with alternative strategies for putting human knowledge on a more secure footing. In effect, the underlying theological anthropology of the Fall thus becomes an excellent predictor of one's preference for a rationalist or an empiricist epistemology.

By and large, rationalists adopted a more optimistic view of the Fall, treating it as a comparatively minor disturbance that left unscathed the "natural light of reason", which continues to provide the basis for the certainty and completeness of human knowledge. Aquinas and Descartes, despite all their metaphysical differences, agree on this important point. The empiricist architects of the Scientific

Method, however, had a far more pessimistic outlook, influenced by Augustinian views about the depth by which human nature had been corrupted as a result of the Fall. Compared to their rationalist brethren, empiricists were highly skeptical about the reliability of human reason, and questioned or denied the possibility of *a priori* knowledge. Experimental modern science thus arose out of a general awareness that an increase in human knowledge in our present condition as fallen creatures – to the extent that this is possible at all – would have to be painstakingly earned through hard labor, such as the trials and tribulations proscribed by Francis Bacon's method of inductive experimentalism. Adopting a thoroughly pessimistic assessment of our post-lapsarian predicament, Bacon was adamant that the instauration of experimental science as the new method of learning could not succeed without the artificial imposition of external constraints: "rigorous testing of knowledge claims, repeated experiments, communal witnessing, the gradual accumulation of 'histories', the use of artificial instruments to amplify the dim powers of the senses, and the corporate rather than individual production of knowledge" (Harrison 2007, p. 51).

In alignment with the more skeptical empiricist line of thought, Clark has espoused a similarly pessimistic account of the "natural" powers of the human mind – understood in the present context as the preciously limited cognitive capabilities Clark deems to be endemic to our biological brains. Certainly, Clark's recurring emphasis on the "messy," "sluggish," and "computationally cheap" cognitive architecture of our biological brains, dutifully hardwired by evolution to serve the exigencies of embodied cognition, does not underwrite the characteristic confidence which rationalist – including Fuller – would want to place in the faculty of human reason. Instead, Clark's situated human cognizer finds herself in the same boat with the Baconian experimental scientist. Encumbered by their evolutionary animal heritage, we can expect our brains to cope reasonably well with the mundane tasks of sensorimotor coordination, associative reasoning, or multimodal sensory integration. Yet, our brains are certifiably baffled by the extraordinary requirements imposed by the demands of human reason, such as deductive reasoning, mathematics, long-term planning, or philosophical reflection. We are, as Clark delights in quipping, "good at Frisbee, bad at logic" (2013, p. 168).

But even in this far from Edenic state of our feeble brain power, the image of our divine likeness hasn't been entirely debased. There is hope for corrective self-improvement, if only in Baconian piecemeal fashion, provided that we scaffold our fallen minds/brains with the right kind of surround, and boost its performance by means of cognitive technologies. As we shall see in Clark's examples below, the most potent forms of cognitive restoration and expansion come from harnessing environmental resources which afford operations that are *complementary* to those which come naturally to our brains. Extending ourselves with staggering layers of self-designed environments thus represents humanity's collective attempt at rendering our fallen minds fit for doing what our biological brains are not – or no longer, if Adamic perfection is any guide – naturally capable of doing. For Clark, much like for Bacon, humanity's best effort at our renewal of Adamic knowledge would be the edifice of human science, a virtuosic display of artificially regimented cognitive scaffolding (Clark 2008, Chapter 3; Clark 2014, p. 172).

In sum, I contend that the empiricist outlook of Clark's EMT is thoroughly "Baconian" for two main reasons: first, because of his pessimistic assessment of the "natural" cognitive endowment of situated human cognizers; second, for his utilitarian reliance on the redemptive qualities of cognitive technology, including the regimes of education through which we must learn to deploy it skillfully. To connect this historical comparison with the main thesis of this section, I now proceed to consider how extended minds (*sensu* Clark) still manage, in their scrappy ways, to at least partially live up to the lofty epistemic ideals that can be derived from the two early modern models of *imago dei*. First, what remnants of the "ideal insight" model can be identified in Clark's EMT?

Many of Clark's most evocative examples of extended cognition rest on the powerful dynamics of "active (cognitive) dovetailing" (Clark 2008, p. 73, 2014, p. 180). By this notion, Clark refers to the outcome of a co-developmental process wherein the neural resources of our malleable biological brains become highly attuned and deeply integrated with external cognitive technologies in the performance of various cognitive tasks. For example, the skilled bartender who learns to rely on differently shaped glasses as a way of recalling incoming drink orders, effectively reducing her inner memory load; the creative artist who overcomes her imaginative limitations through the iterated process of externalizing and re-perceiving her half-baked ideas on a sketchpad; the use of symbolic language as an "augmented reality trick" to overlay and project abstract and otherwise hidden regularities onto the unruly world of sensory perception; and the repurposing of inner speech and outer writing to reflect on one's thought processes, a special kind of "second-order discourse" supporting "a cluster of powerful capacities involving self-evaluation, self-criticism, and finely honed remedial responses" such as recognizing a flaw in one's argument, and fixing it (for a discussion of these examples, see Clark 2008, Chapters 3–4, 2014, Chapter 8). In each of these configurations, inner and outer resources *complement* each other, with respect to the cognitive operations they afford, in ways that fit together as tightly as the sides of a precisely dovetailed joint. By virtue of their coalescence at multiple timescales, and levels of processing and organization, their dovetailing yields cognitive benefits that neither biological nor technological joints alone would be able to support.

The "ideal insight" model of epistemic perfection is directly inscribed into Clark's characterization of the benefits that can be derived from extending one's mind. In his analysis, Clark envisions how our cognitive apparatus would perform if it consisted entirely of parts that are perfectly designed and put together for the task at hand. In doing so, he resorts to the same benchmarks of perfection that were also attractive to early modern philosophers. For example, the perfect bartender would have total recall of the sequence and content of incoming drink orders at all times, serving her customers with utmost confidence that her memories are fully accurate. A supremely creative artist would be able to instantaneously translate her intuitive flashes of brilliance onto the canvas, entirely unconstrained by imaginative barriers such as the "functional fixedness" (Duncker 1945) which limit our ability to shake off perceptual habits. A perfect intellect would have no use for our symbolic creations of logic or mathematics. As the logical empiricist Hans Hahn (1933)

argued, their practical usefulness and ability to occasionally "surprise" us depends entirely on the limitations of human reason. An omniscient agent, to whom the definitions of all fundamental logico-mathematical concepts were fully transparent, would be able to apprehend at a glance the infinite number of propositions that are implied by the former; thus rendering obsolete the need for constructing proofs. Herbert Simon effectively took up this point with his suggestion that "solving a problem means representing it so as to make the solution transparent" (Simon 1981, p. 153). While Simon was primarily concerned with the development of machine intelligence, he saw the importance of creating the right format of symbolic representations for the purpose of automated theorem-proving; unlike Clark, though, Simon didn't fully acknowledge the vicissitudes of situated cognition, nor did he pay the same attention to the ecological subtleties of cognitive dovetailing (Kirsh 2009). Finally, an epistemically ideal cognizer would be in a position to form infallible judgments in all circumstances, perfectly structure her plans, and fully appreciate the logical transitions of her arguments. Thus, she would have no need for linguistically mediated "remedial responses" of the sort envisioned by Clark.

In sum, we can say that Clark has taken aboard the rationalist version of the "ideal insight" model, but, twisted by a "strange inversion of reasoning", supplied it with an empiricist foundation. Recall that for empiricists, the archetypical model of an "ideal insight" is the immediacy and transparency by which we grasp the content of our minds through conscious introspection. During the seventeenth-century "heyday of ideas" (Hacking 1975), empiricists took ideas to be the basic vehicles of cognition, which in many ways they likened to inner images. Thus, when we survey our minds, e.g., to tell whether we have a certain idea or not, we introspect our ideas through a faculty very much akin to sight – the "mind's eye". By redrawing the boundaries of cognition, Clark is able to claim for our dealings with incorporated cognitive technologies a comparable sense of epistemic transparency and infallibility which, for classical empiricists, separates knowledge of our own minds from knowledge of the external world. That is, if our epistemic contact with cognitive artifacts are literally seen as extended realizations of our cognitive apparatus, then what classical empiricists would have considered as epistemically uncertain acts of an inner mind *perceiving* the outer world are functionally akin to epistemically secure acts of *introspecting* one's extended mind. To secure this conclusion, Clark and Chalmers (1998, p. 17) posited a variety of "coupling conditions" that a bio-external resource must satisfy in order to count as part of an agent's extended mind. For example, they required that the external resource should be reliably available and typically invoked on demand, easily accessible, and whatever information it provides should be considered trustworthy and thus more or less automatically endorsed (see Clark 2008, Chapter 5). In essence, the gist of Clark's "trust-and-glue" conditions is to re-cast the empiricist ideal of first-personal epistemic access in functional and informational terms, as a privileged kind of cognitive relationship that situated human cognizers can emulate by actively incorporating parts of their material surround.

Fast-forwarding to present-day technologies, the drive towards expanding the scope and bandwidth of introspection-like self-tracking is conspicuously

exemplified in the 'quantified self' (QS) movement (Swan 2013; Bode and Christensen 2015). Its practitioners embrace an experimental ethos of scrutinizing about every facet of their physiological (e.g., sleep quality), behavioral (e.g., dietary), but also psychological (e.g., mood) and cognitive (e.g., alertness) selves, with the help of specifically designed tools such as biosensors, wearable computers, cell phone applications, and genomic testing kits. While this may strike some readers as a particularly materialistic rendering of the Apollonian maxim to "Know thyself!", dedicated self-trackers conceive their project as a technologically proactive continuation of the ancient desire to improve the quality of one's life through a better understanding of body and mind, i.e., an Epicurean "care of the self" (in the words of the late Foucault). Clark's own illustrations of "behavioral self-scaffolding" are relatively mundane, but neatly demonstrate the superiority of externalized control loops over purely inner thought processes (Clark 2008, Chapter 3.3). For example, the self-directed rehearsal of rule-like mantras (verbalized overtly or in inner speech) aids novices to perform complex pre-planned behaviors, and expert practitioners are skilled at using "instructional nudges" (Sutton 2007) to selectively modulate their attentional and also motivational resources, so they can reliably perform at high levels under pressure (Sutton 2007) Such local routines of behavioral self-scaffolding share with the most engulfing forms of digitally automated self-tracking an adherence to the remedial script of the "ideal insight" model. In the end – to borrow a slogan from Gary Wolf (2009) – devotees of the QS movement are seeking "self-knowledge through numbers" (cited after Bode and Kristensen 2015). That is, wary of the partiality, blind spots and habitual biases of introspection which prevent fallible human decision-makers from making the right choices, they aim to circumvent these deficiencies by resorting to the numerically precise, objectifiable feedback of self-tracking technologies, which they deem as more reliable and trustworthy than the vaguer, and subjectively colored corrective advice given out by professional psychotherapists.

Mutatis mutandis, the same logic helps explain the manner in which remnants of the (materialistically inflected) early modern "ideal agency" model resurface in Clark's EMT. The concrete interpretive lens on offer here is to see our human knack for cognitive extension as a specific development of Simon's broader notion of "bounded rationality", which – as we have seen – Fuller adopted in his reformulated Leibnizian version of the "ideal agency" model of *imago dei*. Simon argued that the rationality of human decision-makers is limited by the amount of information they have at their disposal, the cognitive limitations of their minds, and the finite amount of time they have to arrive at a decision. In Clark's EMT, a generalized version of Simon's notion is taken to govern the entirety of human cognition, sensorimotor processing, and embodied action. Perhaps the best way to illustrate the overarching importance of this principle is through the notion of an *epistemic action*, which Clark and Chalmers (1998) borrow from a classic study of human problem-solving by Kirsh and Maglio (1994).

In their seminal study, Kirsh and Maglio investigated the rationality of human problem-solving in the specific context of the old video game Tetris. To succeed in Tetris, a player has to determine, under severe time pressure, whether

two-dimensional blocks of various shapes, displayed in falling motion on a computer screen, fit into slots in an emerging wall at the bottom of the screen. Classical models of rationality would predict that expert players are better than novices in this game because they are better at performing only actions which bring them closer to the desired goal state, which (in Tetris) is to maximize the surface area of the emerging wall. This makes sense from an intuitive standpoint, under the natural assumption that rationality consists in being able to navigate some problem-space quickly and efficiently. For example, we'd expect that experts in spatial navigation should be able to find a shorter path out of a maze. However, Kirsh and Maglio found that expert players of Tetris failed to conform to this expectation. They often manipulated the falling tiles in ways that would seem to be disadvantageous and indeed "irrational" from a traditional perspective. For instance, experts would often rotate blocks on top the screen just to see whether they fit into one of the open slots. The key point to notice here is that if it turns out that the rotated block doesn't fit, then performing this type of action has not brought the player any closer to her goal. At best, the action would be gratuitous; at worst, it may even lead the expert further away from the goal state. Interestingly, the same type of seemingly "erratic" behavior occurred much less frequently among novice players. Having to decide where to put a falling tile, novice players usually did not perform any overt actions at all, but attended entirely to mental images of the falling blocks which they rotated in their heads. *Prima facie,* this behavior contradicts the intuitive assumption that expert players of Tetris are more rational problem-solvers than are novices.

In their analysis, Kirsh and Maglio found an ingenious way of explaining away this apparent paradox, by distinguishing between *pragmatic* and *epistemic* actions. While the former are actions designed to bring an agent physically closer to the goal state, such as finding one's way out of a maze, epistemic actions are designed to improve one's informational state as a problem-solver, such as uncovering information that would otherwise be hidden. Doing so may or may not physically advance an agent to the desired goal state, but nevertheless pays off if the action yields epistemic dividends which outstrip its physical costs. This is precisely what the Tetris experts became so prolific in doing – they learned to maximize the benefits of *epistemic agency.* As Kirsh & Maglio show, rotating tiles directly on the screen is a faster and more reliable strategy than rotating mental images of tiles in one's head. Thus, expert players of Tetris were indeed more rational, after all – provided that we analyze the costs and benefits of their actions as embedded in a joint problem-space that includes both physical and epistemic goals (cf. Clark 2008, Chapter 9). Kirsh and Maglio offer several descriptions of the generic cognitive benefits that people can accrue from epistemic actions – "actions that use the world to improve cognition," "actions [that] are used to change the world in order to simplify the problem-solving task," and "actions performed to uncover information that is hidden or hard to compute mentally" (Kirsh and Maglio 1994, p. 2). Glossing over these subtle differences, the upshot is that epistemic actions reap cognitive benefits because they reduce cognitive load, simplify the path to a correct solution, reduce the likelihood of error, or some combination of the above. Once we start looking, epistemic actions are found everywhere: we move our hands close to something hot before grasping

it; we squint our eyes to see a distant object more clearly; carpenters know the value of measure twice, cut once; and savvy readers underline important passages (Kirsh 2006, p. 252).

More generally, at its core, the bounded rationality of situated human problem-solving requires the solution of constant trade-offs between pragmatic and epistemic actions. The exact amount of these trade-offs will depend on the precise nature of the cost function, which is determined by many, in part conflicting considerations and constraints (Kirsh 2006, 2013). But in principle, an optimal solution can be calculated by the mind of a situated human cognizer, i.e., a creature equipped with a less-than-perfectly rational Leibnizian will (*sensu* Fuller). From this standpoint, the necessary embodiment of situated human cognition turns out to be an asset rather than a liability, because it is only by *acting in the world* that we can drastically reduce the total cost of the actions and cogitations we choose to perform. Consequently, Clark frequently appeals to "minimal cognitive effort" principles to highlight the relevance of embodiment for mind and cognition (e.g., Clark 2008, Chapter 9.1). Those include the human potential for "spreading the load", by exploiting the morphology, actions, or biomechanics of one's body, but also through environmental structure and interventions; the "self-structuring of information" by eliciting good data, and active learning strategies in general; and the myriad ways in which we extend our minds through the incorporation of cognitive technologies (Clark 2008, p. 196–97). In Clark's hands, the trend towards "ephemeralization", which Buckminster Fuller keenly observed in the history of technological progress, is thus revealed to be a governing principle of extended human cognition. As we shall discuss further in the next section, Clark's reliance on epistemic agency as a primary vehicle of extended cognition is prefigured in the anthropology of Marx and Engels, in particular their emphasis on *labor* as the fundamental source of human value; although it will not have escaped the reader's attention that in Clark's stock examples, manual (pragmatic) and intellectual (epistemic) labor are integrated rather harmoniously in the service of individual cognitive success.

13.4 Self-Made (Cyborg) Man Aspiring to the "Practice Ideal" of Imago Dei

When I first introduced Clark's portrayal of human beings as *natural-born cyborgs* (sect. 13.1), I noted how the EMT radicalizes the familiar trope of "man the tool-maker". With his cyborg vision of the human mind, Clark has vividly emphasized that we have always been good at adapting our minds, skills, and practices to the demands of our tools and artifacts, and *vice versa*; a reciprocal process Clark describes as "active cognitive dovetailing" (Clark 2003, 2008; see above). Propelled upward by "a virtuous spiral of brain/culture influence" (Clark 2013, 180), boundless opportunities beckon for natural-born cyborgs – a momentous chain of events

made possible by the most marvelous of human organs, the brain, which Clark has aptly depicted as "nature's great mental chameleon" (Clark 2003, p. 197).

Indeed, it is hard to overstate the importance which Clark attaches to his view that "plasticity and multiplicity are our true constants" (Clark 2003, p. 8) – phrases which reverberate throughout his writings. For Clark, it seems a given that our necessary openness to information-processing mergers and coalitions – the polymorphous nature of our "soft selves" – lies at the very heart of what makes us human. Therein, I discern another deep affinity between Clark's EMT and the theological scaffolding of Fuller's transhumanist project. In this final section of the paper, I thus elaborate on the intellectual history of the prominent leitmotif of human plasticity and openness, by revealing its roots in yet another historically consequential rendering of *imago dei*, the so-called "practice" model, which also links EMT to the dialectical-materialist anthropology of Marx and Engels (cf. Craig 1987, Chapter 5).

What is widely regarded to set Clark's EMT apart from more moderate conceptions of human cognition as 'scaffolded' or 'mediated' by an ever-expanding toolkit is the ontological heft it attaches to the causal entanglements of our brains and bodies with the environment, culture, and technology (Malafouris 2013; Theiner and Drain 2017). As we have seen, Clark has vigorously argued for cognitive extension in a number of ways we shall not rehearse here (Clark 2008; Theiner 2011). Instead, I would like to foreground here the fact that all of Clark's arguments pivot on criteria for individuating minds that are not biologically but functionally based (e.g., the depth, reliability, and interactional complexity of human-artifact "couplings"; cf. Clark 2008; Heersmink 2015; Clowes 2015). In effect, these criteria shift what is most valued about being human – at least cognitively speaking – from our contingent embodiment in an organic body to whatever mental capabilities those bodies afford, perhaps only in a woefully underdeveloped form that stands in need of further promotion. Defending the EMT, Clark has openly expressed his preference for a (suitably nuanced) functionalist understanding of mind as "a flexible and information-sensitive control system for a being capable of reasoning, of feeling, and of experiencing the world (a 'sentient informavore' if you will)" (Clark 2008). Within the larger cognitive assemblies in which we, as sentient informavores, think and act, Clark notably assigns a special role to the human brain, as the powerful "master orchestrator" of the myriad ways in which the cyborgian human mind can extend itself (Clark 2008, p. 139). This point, I shall now argue, betrays a deeper ambivalence within Clark's writings concerning the role of the brain, considered as humanity's most distinctive biological organ (Fuller 2013, Chapter 5). For all his emphasis on embodied prediction and situated action (Clark 2008, 2015), my hunch is that the real reason why Clark continues to privilege the brain over the (non-neural) body is because he sees it as the "ultimate originator" of our cognitive self-transcendence.[13] What makes the human brain special, for Clark, is that it initially set off, but also continually sustains the snowballing process of enculturation whose

[13] Much to the chagrin of Hutchins (2011), who takes Clark to task for his relative neglect of the role played by cultural practices.

long-term effects on the mind are to enable precisely those distinctively human "top-level" cognitive achievements through which we exercise – as Fuller would say – our most God-like capacities.

If I my hunch is right, it would thus be a mistake, as this is sometimes done, to subsume Clark's EMT squarely under the banner of "embodied cognition", a family of views which stress the deep, and in some ways unexpected dependence of mind and cognition on the fine-grained details of embodiment (see Wilson and Foglia 2017 for an overview). From Clark's EMT perspective, the "ancient biological skin-bag," as Clark rather disparagingly refers to the body, is primarily a "handy container of persisting recruitment processes and of a batch of core data, information, and body-involving skills; thusly equipped the mobile human organism is revealed as a kind of walking BIOS" – in short, a fungible platform which allows extendible minds to continually alter and re-program their cognitive profiles (Clark 2008, p, 138). In similar vein, when Clark draws his distinction between "modest" and "profound" grades of embodiment, he stresses the functional *open-endedness* and morphological *plasticity* as fundamental to human minds (Clark 2008, pp. 42–43). In contrast to mind-body dualism, the freedom of a "profoundly embodied" agent is, of course, not that of an immaterial spirit who fancies itself outside the deterministic nexus of the material world, but that of a deeply *promiscuous* creature freed from the shackles of having "a fixed mind (one constituted solely by a given biological brain) and […] a fixed bodily presence in a wider world" (Clark 2008, pp. 42–43). Here, transhumanist would urge Clark to go one step further, and consider biological embodiment as such as a transitory confinement through which humanity will eventually bootstrap its way from our shared animal heritage into radically novel evolutionary arrangements, which they deem better suited to realize a fuller spectrum of human potential.

Let me bring out Clark's ambivalence about the status of the human brain more forcefully by casting it in terms of Fuller's (2013, Chapter 5.3) contrast between "organic" versus "machinic" conceptions of human nature, and their respective viewpoints regarding the relationship between mind and life (Fuller 2013, Chapter 5.3, 2014, 2018). For the organic conception, mind and life are seen as deeply continuous. Hence, the brain shares an important kinship with all other biological organs, in virtue of providing the necessary psychological organization and thus behavioral flexibility to move around the body of an upright ape. Earlier in his opus, Clark frequently stressed this point: "Biological brains are first and foremost the control systems for biological bodies. Biological bodies move and act in rich real-world surroundings" (Clark 1998, p. 506). In order to play this role effectively, Clark then argued, human cognition must by design be deeply embodied, situated, and geared towards effective action (cf. Wilson 2002). From an organic perspective, the primary ontological affiliation of humankind would lie with similarly embodied creatures within the animal kingdom.

From a machinic perspective, however, the continuity between mind and life is tenuous at best, and may indeed become entirely obsolete if we find ways of migrating our minds from their current carbon-based containers to more hospitable digital environments, which would presumably be implemented in silicon-based

computers. Within analytic philosophy of mind, such a perspective is closely aligned in spirit with the viewpoint of "machine functionalism" made popular by Hilary Putnam in a couple of seminal articles in the 1960s (Putnam 1960, 1967). Drawing on the abstract notion of *computation* as developed by Turing and others, Putnam argued that we ought to resist the blatantly "species-chauvinistic" identification of mental states with the specific kinds of neural states that we find in human (or mammalian) brains, on roughly the same grounds on which we take software to be distinct from hardware.[14] Putnam's functionalism was readily embraced by the physicalist orthodoxy, and grudgingly endorsed by the dualist minority, for its decidedly anti-reductionist stance; besides, it conveniently provided a much-needed philosophical foundation for classical AI research and the nascent field of cognitive science. Similarly, by taking a machinic perspective on the role of the brain, Fuller means to stress the "formal similarities between ourselves and other creatures who may be quite differently embodied and even normally operate on a different plane of reality (e.g. God, angels, machines)" (Fuller 2013, p. 97). From this perspective, our primary affiliation would thus lie with creatures who reign supreme in cognitive domains that we value highly – irrespective of their embodiment, and regardless of whether they can be considered biologically alive.

Notably, Fuller's allusion to "machines" doesn't surrender us to the fate of determinism, in the sense in which (e.g.) Descartes would have subsumed material things under the realm of "mechanism", as opposed to the realm of freedom occupied by immaterial souls. On the contrary, it is meant to reflect back on the creative powers of their human makers, and the traditional association of human technological ingenuity with our having been created *in imago dei* (see Noble 1997, esp. Chapter 10) – a metaphysical impulse that might also explain why transhumanists are enamored of the prospect of creating super-intelligent machines (Bostrom 2014). More generally, the machinic conception accords the human brain a special status vis-à-vis all other organs, as the premier site of cognitive transcendence. This means that unlike the function of any other "merely" life-supporting biological organ, the unique cognitive function of the human brain is to enable an empowering web of artifacts, technology, and culture, allowing it (in controlled ways) to reshape who we are, and thus who we are capable of being. I contend that Clark's sympathy for a "machinic" conception of mind lies behind his lingering attachment to the "hypothesis of organism-centered cognition," with its concomitant privileging of the human brain (Clark 2008, p. 139). On the machinic conception, then, the ultimate purpose of the brain is not so much to deal with the vicissitudes of embodied cognition, but to help our minds break free from a biology-bound heritage that is dictated by the evolutionary lineage of our animal bodies.

Lest we chide Clark and transhumanists alike for the "techno-philic" orientation of their viewpoints, it will be helpful to place their fondness for the machinic aspects of human existence into a different historical context. In his treatise on *Creative*

[14] On the impact of cybernetics on twentieth-century developments of a de-physicalized notion of 'machine', including Alan Turing's own change of mind regarding the prospect of creating 'machine intelligence', see Malapi-Nelson (2017, esp. Chap. 4).

Evolution, Henri Bergson famously suggested to redefine the human species as "not *Homo sapiens*, but *Homo faber.* In short, *intelligence, considered in what seem to be its original feature, is the faculty of manufacturing artificial objects, especially tools to make tools, and of indefinitely varying the manufacture*" (Bergson 1998/1911, p. 139). To be sure, Bergson's facile pronouncement that technology is a unique display of human creativity has been defeated by the discovery of a great variety of tool-making and tool-using abilities that are surprisingly common among non-human animals (Shew 2017). Still, what unites Clark, Bergson, and consorts in their emphasis on technology as the foundation of becoming-human ("anthropogenesis") is their focus on the *higher-order* ability of humans to produce tools for making tools, the sheer open-endedness of cultural techniques we devise for using these tools, and the complexity of artifacts we are capable of producing with their help (Stiegler 1998; Malafouris 2013). More precisely, for Clark and his allies, the discontinuity in kind which they posit between human technology and animal tool use is grounded in the endlessly "looping" nature by which our minds, brains, bodies, are entwined with material culture. In this vein, the EMT-friendly cognitive archeologist Lambros Malafouris has proposed that "*meta-plasticity* – the fact that we have a plastic mind which is embedded and inextricably enfolded with a plastic culture – might well be the locus of human uniqueness *par excellence* (Malafouris 2013, p. 46, Epilogue).

To be sure, transhumanists will look at the long-term opportunities which our meta-plasticity affords primarily from an enhancement angle. With this minor perspective shift, we can rephrase the Bergsonian slogan of *homo sapiens* as *homo faber* accordingly, by calling us the first "recursively self-improving" species on an evolutionary timescale.[15] In AI circles, the notion of "recursive self-improvement" has been widely discussed as the ability of a "seed AI" with at least human levels of general-purpose intelligence to iteratively enhance itself, which carries in itself the eventual possibility of an uncontrollable intelligence explosion (a "singularity"; cf. Bostrom 2014). When viewed from the standpoint of good-old-fashioned Darwinian evolution, the seed of such an intelligence explosion was arguably already sown when *homo sapiens* arrived on the scene, or at least once humanity gained the requisite level of insight and control over nature to alter and redirect the course of its future evolution. As mentioned earlier (Sect. 13.1), it was the fact that the evolution of biological life had become "reflexive" with our discovery of DNA, and the prospects afforded by genetic engineering, which originally prompted Julian Huxley (1957) to coin the term "transhumanism". Less dramatically, Clark's conception of human beings as "profoundly embodied" agents similarly foregrounds the cumulative complexity of our cognitive self-improvement efforts, noting in particular the role played by educational practices and institutions:

> "We do not just self-engineer better worlds to think in. We self-engineer ourselves to think and perform better in the worlds we find ourselves in. We self-engineer worlds in which to

[15] But presumably not the last, if we contemplate the creation of super-intelligent machines, or the coming of posthuman beings as a potential successor species.

> build better worlds to think in. We build better tools to think with and use these very tools to discover still better tools to think with. We tune the way we use these tools by building educational practices to train ourselves to use our best cognitive tools better. We even tune the way we tune the way we use our best cognitive tools by devising environments that help build better environments for educating ourselves in the use of our own cognitive tools (e.g., environments geared toward teacher education and training). Our mature mental routines are not merely self-engineered: They are massively, overwhelmingly, almost *unimaginably* self-engineered." (Clark 2008, pp. 59–60).

Although the suggested analogy between the neural effects of enculturation, the formation of educational ideals and practices, and the effects of technology-driven cognitive enhancements remains contested (see Harrison 2010; Buchanan 2011), it is broadly supported by the framework of the EMT. If we are, at bottom, a self-made "cyborg" species, then the power and inherent drive to transform ourselves in ways we deem desirable, prudent, and morally praiseworthy, has always been a genuine part of what makes us human. In his treatment of *Humanity 2.0*, Fuller notes the "inherent artificiality" of the human by dubbing the quest for "spiritual" or "intellectual" growth as a "call to artifice" (2011, p. 74). Fuller's assessment is based on the supposition that among the diverse set of traits which, as a biologically contingent matter of fact, make up the conditions of our organic existence, not all may be equally conducive to the fullest expression of human potential. It follows that our general level of humanity may be elevated by strategically interfering with the supposedly "natural order" of things. Historically, the "call to artifice" has motivated a great variety of "humanist" projects and disciplines, including policies for promoting more equitable distributions of wealth, or the provision of mass education and welfare by the state (see also Hughes 2004).

Within the Western tradition, Fuller has identified three main "ages of artifice" that trade on variations of this more general idea: (1) *the ancient artifice*, exemplified in the Greek ideal of *paideia*. It promoted the teaching of a range of cognitive skills and practices (the "liberal arts") that were considered essential for playing an active part in civic life ("worthy of a free person", hence far from inclusive); (2) *the medieval artifice,* exemplified by the introduction of the legal category of a *universitas*, i.e., an artificial corporate person, into Roman law (Kantorowicz 1957/2016; see also Fuller 2011, p. 104). Its legal construction enabled the literal "incorporation" of human beings for the pursuit of collective ends which far exceed the lifetimes and capabilities of individuals, and – unlike families, dynasties, or clans – without being tied to biological categories such as kinship or reproduction; (3) *the modern artifice,* exemplified by the emergence of engineering as a distinct profession, which is based on the creative application of scientific knowledge for redesigning the natural world, including our biological bodies. With his invocation of the "inherent artificiality" of humanity, Fuller thus effectively agrees with Foucault's demystified account of "humanity" as a historically contingent social construction, but takes a different tact, by fixating precisely on the "ontological precariousness" of the human as a way of motivating his adoption of a transhumanist stance.

Here it is worth stressing, though, that absent a substantive account of the normative horizons towards which transhumanists are eager to strive, a "call to artifice" alone offers no guarantee to think their journey is worth embarking (Hughes 2004; Bostrom 2008; Fuller 2011, 2013). Granting the transhumanist premise that technological engagement can fulfill a vital role in the acceleration of human progress, the aforementioned "pedagogical, corporate, and techno-scientific" vehicles of enhancement need not (and do not) always coalesce in ways that support the betterment of humanity. McLuhan (1964, p. 55) famously foresaw the possibility of a treacherous turn, that "[b]y continuously embracing technologies, we relate ourselves to them as servo-mechanisms". Indeed, his envisioned reversal between means and ends has become a real danger as current information and communication technologies have become increasingly autonomous, and even imbued with their own agency: i.e., they actively, automatically, continually, seamlessly, flexibly, and unwittingly reshape themselves not only to fit pre-existing human needs and ends, but to actively shape and alter our cognitive and behavioral profiles as a means to others' ends (Smart et al. 2017). In this artificially ramped-up state of "cognitive dovetailing", there is no longer a rigid boundary between mind and world, or between person and tool, nor is there a clear separation between tools and the environments in which they are embedded (Aydin et al. 2019). Hence, McLuhan's worry that humans may end up trapping themselves in the technologies of their own creation must be seen as a matter of real concern. Social critics have decried our obsessive, ever-intensifying absorption with digital technologies as the source of "distracted minds" (Gazzaley and Rosen 2016), and warn that the flip side of cognitive extension is "mind invasion" (Slaby 2016), more often than not driven by corporate interests to further the engineering of "predictable and programmable people" (Frischmann and Selinger 2018). Despite these caveats, Clark, much like Fuller, has remained guardedly optimistic in his assessment that such problems are, in the end, "foreseeable but unintended" side effects, local pathologies in a digitally networked cognitive ecology that can and need to be dealt with in the specific contexts in which they occur (Clark 2003).[16] Be that as it may, a blind trust in the capabilities of techno-social engineering to deliver humanity's greatest aspirations would seem to be ill-advised.

After this caveat, let us return to the contrast between organic and machinic conceptions of the human mind that we sketched. Another way of framing this contrast would be to ask – echoing the question posed by modernist movements in twentieth century art – *whether function follows form* or *form follows function* when it comes to the question of what it means to be human (see Fuller 2013, Chapter 3.2). Today, as the once-impermeable ontological divide between animals and humans on the one hand, and humans and machines on the other, is becoming increasingly porous, questions about the future identity of the human, but also related questions about animal and machine rights have moved to the forefront of public discourse. While

[16] I allude here to the familiar "doctrine of double effect" – i.e., a principle that has often been invoked to justify the moral permissibility of an action that causes serious harm, as an unintended *side effect* (though not a means) of promoting some good end.

posthumanists stress the existence of all-encompassing communal bonds that unite humanity with all forms of life (or matter in general), such as the "zoë-egalitarian" viewpoint espoused by Braidotti (2013), transhumanists stress the contingency of our association with animal bodies by pointing to our desire for entering increasingly "virtual" modes of existence. As Fuller (2013) points out, this includes our obsession with digital media, our readiness to immerse ourselves into virtual realities, and the tendency of many people to identify more easily with their digital avatars than with their own bodies (and, one might add, more so than with their pets or their simian ancestors).

In Clark's work, these competing conceptions of what we most cherish about being human play out as an explanatory contrast between "modest" versus "profound" conceptions of embodiment (Clark 2008, Chapter 2.7). Instances of the former, which Clark collects under the heading of the "active body" (Clark 2008, Chapter 1), include the "passive-dynamic walker" whose natural-looking gait exploits the kinematics and organization of its mechanical linkages and other components; the efficient devolution of neural control to one's body (aka "morphological computation"), or the use of repeated eye fixations as "deictic pointers" in the aid of perceptual processing. Common to these clever displays of "modest" embodiment is a general recognition of "the important contributions that embodiment and environmental embedding can make to the solution of a problem […] in real-time performance of the task" (Clark 2008, p. 14). In these displays, *the functionality of a cognitive process is boosted by getting the most out of the biologically constrained form of one's body.*

In contrast, for "profoundly" embodied agent such as human beings, no such material constraints are ever taken as fixed points, as constants that would inevitably define the problem space in which we are forced to think, act, and conduct our lives. Instead, the minds of natural-born cyborgs are "promiscuously body-and-world exploiting" (Clark 2008, p. 42), continually seeking ways to exchange its working parts, expand its boundaries, and flexibly redefine its interfaces to the wider world. In essence, *the material form of a profoundly embodied agent is in principle negotiable, and tailored to whatever functions the agent fancies herself to perform*; importantly, those functions co-evolve with the technological and socio-cultural environments which we create, but which equally have their say in creating *us* (Clark 2003). Hence, it wouldn't be a stretch to argue, endorsing a machinic perspective, that our biological bodies are nature's own enabling technology through which extended minds manage and control their cognitive mergers and acquisitions, at least to the point where they are ready to escape their biological containers for good. Later in his book, Clark (2008) seems to suggest as much:

> "Finally, the body, by being the immediate locus of willed action, is also the gateway to intelligent offloading. The body…is the primary tool enabling the intelligent use of environmental structure. […] But I am inclined to go further and to assert not just that this is what the body *does* but that this […] is what, at least for all cognitive scientific purposes, the body *is*. I am inclined, that is, to simply identify the body with whatever plays these (and doubtless some additional) roles in the genesis and organization of intelligent behavior." (Clark 2008, pp. 206–07)

I take this passage to be a rather unequivocal approval of a machinic conception of human nature. Notably, it also trades on a fairly traditional "Pelagian" conception of the will as the faculty of human self-transcendence, understood here primarily in the register of cognitive extension. Moreover, by promoting an evaluative assessment of our animal bodies in terms of their functional capabilities (and their limits), the passage also comes close in spirit to an endorsement of a principle for which transhumanists have long been campaigning – the right to "morphological freedom," defined by Sandberg as "an extension of one's right to one's body, not just self-ownership but also the right to modify oneself according to one's desires" (Sandberg 2013, p. 57). Paralleling Clark's distinction between "modest" and "profound" embodiment, Sandberg likewise emphasizes that morphological freedom involves not merely the ("modest") right to exploit the inherent potentialities of its current, *actual* form, but the ("profound") right to the augmentation of human potential through the self-directed modification of *possible* forms of embodiment. As Sandberg puts it, "[h]umans are ends in themselves, but that does not rule out the use of oneself as a tool to achieve oneself" (Sandberg 2013, p. 63).

However, my suggestion that Clark may be flirting with the principle of morphological freedom in the above passage should not be taken to signal a return to the disembodied Cartesian conception of mind as an ethereal *res cogitans*, radically separated from the material side of human existence. In this context, it is worth recalling that in its rejection of Cartesian mind-body dualism, Clark's vision of the cultural domestication of our animal bodies that can only be achieved through the application of cognitive technologies takes up the mantle of a dialectical-materialist anthropology of the sort championed by Marx and Engels. As good materialists, they eagerly sought to discard the "idealist fancy" that our existence derives its meaning and purpose from some other-worldly source, which they loathed for tearing asunder humanity from the rest of the natural world. By viewing this-worldly *labor*, i.e., the deliberate physical reorganization of nature, as the fundamental source of human value, Marx and Engels meant to emphasize that human beings are, and always remain, part of nature; but at the same time, they are unique insofar as they also transcend it. For Marx and Engels, the emancipatory emergence of human agency and consciousness from its material conditions is forged only historically, through the productive powers of labor, which under a capitalist regime (that they viewed as a state of alienation) appears as the creative power of capital.

Their metaphysical conception of the species-becoming of humanity ("anthropogenesis") was given more concrete form in a prescient essay by Engels (1883/1946). Since labor, for Engels, begins with the making of tools, the first critical transition "from ape to man" was the development of bipedalism, because it freed up human hands for the extraction of value through the manual appropriation of natural resources. Engels goes on to propose a speculative developmental sequence of further transformations, such as the evolution of verbal speech, animal husbandry, and control of the fire that successively expanded the ambit of human *praxis*, which in turn spurred the cognitive development of our brains. Central to this co-evolutionary account of anthropogenesis – called "dialectic" in Marxist parlance, though fully consonant with contemporary accounts of "triadic niche construction" (Iriki and

Taoka 2012) – is again the pivotal role that Engels assigns to the powers of the human will. Considering labor and speech as the two key drivers "under the influence of which the brain of the ape gradually changed into that of man[sic], which, *for all its similarity is far larger and more perfect*", Engels emphatically asserts that "But all the planned action of all animals has never resulted in impressing the stamp of their will upon nature. For that, man was required" (Engels 1883/1946, p. 284). Engels' emphasis on our ability to attain conscious control over nature, through an understanding of nature's laws, and our ability to make those laws work towards humanly defined ends, reflects an important intellectual shift that had occurred, over the Romantic era, to the predominant conception of *imago dei* (Craig 1987, Chapters 4–5).

Supplanting the "ideal insight" model which had reigned supreme during the early modern period, in which humans were primarily conceived as passive spectators privy to grasping the immutable order of a reality that was preordained by God, a more active "practice ideal" model took root, which above all things prized our ability of enacting change in the world by *doing*. According to this model, our God-like talents are most clearly exhibited when we impose human form on sensuous material, thereby refashioning our "lifeworld" in ways that are conducive to human flourishing. This reversal of priorities, in which a newly-found reverence for *activity*, *creativity*, and *practice* replaced older contemplative ideals, was facilitated (if not prompted) by two concurrent philosophical developments: first, the reigning conception of a deity to whom human beings are thought to be alike became predominantly pantheistic, conjuring an image of God – Spinoza-style – that does not transcend nature, but is immanent to it; second, as a simile of the ever-changing process that is nature, the static image of a transcendent God, contemplating the eternal truths of his own creation, was replaced by that of a dynamic deity that is engaged in endless self-transformation (Craig 1987, pp. 225–26). In effect, the emerging "practice ideal" conception of *imago dei* implied that we are at our most God-like when we actively transform and create our environments, and, in turn, our undeniable success in doing so could thus be regarded as a vindication of our own divinity. This picture clearly appealed to Marx and Engels' aspiration of making humans masters of their own fate – and thus capable of grasping that the laws of *social* action are under our control, rather than foreign to man (aka "natural"). Echoing Hegel, Engels went so far as to proclaim this historical development of self-consciousness as "the ascent of man from the kingdom of necessity to the kingdom of freedom" (Engels 2015, p. 66).

With his coinage of the "natural-born cyborg," Clark has crafted an updated version of the "practice ideal" model of *imago dei*, by linking the image of ourselves as the "self-made" species to the design, use, and incorporation of cognitive technologies. Thanks to our uniquely extendible minds, we alone are the species that is capable of negotiating the modes of its own embodiment, and thus free to define our relationship to nature and to one another through the profound engagement with artifacts, technologies, and material culture (Malafouris 2013). In a more traditional register, efforts to "enhance" the engagements through which this transformation is taking place would have been spoken of as calls for the "perfectibility" of

humankind (Passmore 1970). Transhumanists evidently continue the "perfectibilist" tradition with their advocacy for the enhancement of the cognitive, emotional, physical, and also moral dimensions of human well-being by means of science and technology (Bostrom 2008; Savulescu and Bostrom 2009). Historically, a prominent antecedent of Clark's "cyborg" ideal of human perfectibility can be found in the movement of Renaissance Humanism, epitomized in its call to locate the source of human dignity in our unique freedom to *choose our own nature.*

In this context, Pico de Mirandola's "Oration on the Dignity of Man" (1468/2016) – are frequently cited as a paragon of transhumanist aspirations[17] (e.g., in Bostrom 2008). In the famous opening pages, Pico speaks of human beings as creatures of a uniquely "indeterminate nature," endowed with a boundless capacity for self-transformation which places us at the summit of creation. Pico accounts for this exceptional status of humanity – our "ontological precariousness" in Fuller's idiom – as God's response to a dilemma He had to overcome afterdeciding to create a rational being capable of contemplating the fullness and beauty of His divine handiwork. God's dilemma was that He had already exhausted every possible nature in process. Thus, seeing as there wasn't any fixed essence left after which human nature might be formed, God proceeded to give humans an indeterminate nature, making it Adam's prerogative to fashion one for himself:

> "Adam, we give you no fixed place to live, no form that is peculiar to you, nor any function that is yours alone. According to your desires and judgment, you will have and possess whatever place to live, whatever form, and whatever functions you yourself choose. All other things have a limited and fixed nature prescribed and bounded by our laws. You, with no limit or no bound, may choose for yourself the limits and bounds of your nature. We have placed you at the world's center so that you may survey everything else in the world. We have made you neither of heavenly nor of earthly stuff, neither mortal nor immortal, so that with free choice and dignity, you may fashion yourself into whatever form you choose. To you is granted the power of degrading yourself into the lower forms of life, the beasts, and to you is granted the power, contained in your intellect and judgment, to be reborn into the higher forms, the divine." (Pico 1468/2016).

It is worth quoting this passage in its entirety, because contrary to its popular reception – air-brushed through post-Kantian history of philosophy – as a secular prelude to modern conceptions of morality as grounded in human freedom and autonomy, it reveals how the anthropocentrism of Renaissance Neo-Platonism in particular was deeply rooted in religious convictions. Their emancipatory concern for the *studia humanitatis* was, first and foremost, an endeavor to define the relationship of the human to the divine, nourished and sustained by a yearning to reunite with God through the pursuit of spiritual excellence (Copenhaver 2016). Looking past the opening passages of the *Oration*, the Neo-Platonist Pico professes that the ultimate purpose of *paideia* – a technique for instilling in humans the power to

[17] As Sorgner (2016, p. 142) helpfully points out, Pico's Neoplatonism makes him a rather unrepresentative figure of Renaissance Humanism. Pico extolled the perfection of the *intellectual* virtues, pursued for the sake of "purification" from one's earthly attachments, as the path towards mystical reunion with God. This is not exactly the ideal of a "well-rounded" person which is frequently associated with Renaissance Humanism.

assume "whatever form we choose" – is the goal to become like angels, thus aspiring to become *more-than*-human. Those are the esoteric portions of Pico's work, in which he commends the use of magic and Kabbalah to achieve the ethereal, disembodied state of the celestial intelligences – a necessary transformation he thought would eventually enable our mystical reunion with God. This idea is rooted in Aquinas' conception of angels, *qua* spiritual creatures, as lacking in prime matter, which in material beings accounts for the individuation of distinct individuals belonging to the same species. Thus, by reaching this most exalted, angelic state, human beings would cease to exist as a plurality of material persons, divided from one another, and thus also brought closer to God. Pico's radically optimistic view of the self-transcending powers of the human mind is rightly seen as an epitome of Renaissance Pelagianism, with its almost complete disregard for the redemptive role which in orthodox Christianity would be reserved for divine grace (Passmore, Chapter 5). Continuing this tradition, we can discern a path to the belief of Enlightenment philosophers and their transhumanist heirs that humankind may be perfectible by natural (as distinct from supernatural) means, through piecemeal scientific progress and the technological appropriation of nature.

References

Adams, M. M. (1999). *What sort of human nature?* Milwaukee: Marquette University Press.

Aydin, C., Woge, M., & Verbeek, P.-P. (2019). Technological environmentality: Conceptualizing technology as a mediating milieu. *Philosophy of Technology, 32*, 321–338.

Benyus, J. (1997). *Biomimicry*. London: Penguin.

Bergson, H. (1998). *Creative evolution*. Mineola: Dover Publications.

Bode, M., & Kristensen, D. B. (2015). The digital doppelgänger within. A study on self-tracking and the quantified self movement. In R. Canniford (Ed.), *Assembling consumption: Researching actors, networks and markets* (pp. 119–134). London: Routledge.

Bostrom, N. (2005). The history of transhumanist thought. *Journal of Evolution and Technology, 14*(1), 1–25.

Bostrom, N. (2008). Why I want to be a posthuman when I grow up. In B. Gordijn & R. Chadwick (Eds.), *Medical enhancement and posthumanity* (pp. 107–136). Dordrecht: Springer.

Bostrom, N. (2014). *Superintelligence*. Oxford: Oxford University Press.

Bostrom, N., & Sandberg, A. (2009). Cognitive enhancement: Methods, ethics, regulatory challenges. *Science and Engineering Ethics, 15*(3), 311–341.

Braidotti, R. (2013). *The posthuman*. Cambridge: Polity Press.

Brey, P. (2000). Theories of technology as extension of human faculties. In C. Mitcham (Ed.), *Metaphysics, epistemology and technology* (pp. 59–78). London: Elsevier/JAI Press.

Brooke, J. H. (2005). Visions of perfectibility. *Journal of Evolution and Technology, 14*(2), 1–12.

Buchanan, A. E. (2011). *Beyond humanity?* Oxford: Oxford University Press.

Burdett. (2011). Contextualizing a Christian perspective on transcendence and human enhancement*:* Francis Bacon*,* N. F. Fedorov*,* and Pierre Teilhard de Chardin. In R. Cole-Turner (Ed.), *Transhumanism and transcendence* (pp. 19–35). Washington, DC: Georgetown University Press.

Cabrera, L. Y. (2015). *Rethinking human enhancement*. New York: Palgrave Macmillan.

Catholic Church. (2012). *Catechism of the Catholic Church*. Vatican City: Libreria Editrice Vaticana.

Clark, A. (1998). *Being there*. Cambridge, MA: MIT Press.
Clark, A. (2001). Reasons, robots and the extended mind. *Mind & Language, 16*(2), 121–145.
Clark, A. (2003). *Natural-born cyborgs*. New York: Oxford University Press.
Clark, A. (2007). Re-inventing ourselves: The plasticity of embodiment, sensing, and mind. *The Journal of Medicine and Philosophy, 32*(3), 263–282.
Clark, A. (2008). *Supersizing the mind*. New York: Oxford University Press.
Clark, A. (2013). *Mindware* (2nd ed.). New York: Oxford University Press.
Clark, A. (2015). *Surfing uncertainty*. Oxford: Oxford University Press.
Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7–19.
Clowes, R. (2015). Thinking in the cloud: The cognitive incorporation of cloud-based technology. *Philosophy and Technology, 28*(2), 261–296.
Cole, M. (1996). *Cultural psychology*. Cambridge, MA: Belknap Press.
Cole-Turner, R. (Ed.). (2011). *Transhumanism and transcendence*. Washington, DC: Georgetown University Press.
Copenhaver, B. (2016). Giovanni Pico della Mirandola. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2016). Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/fall2016/entries/pico-della-mirandola/
Craig, E. (1987). *The mind of God and the works of man*. Oxford: Clarendon Press.
de Grey, A., & Rae, M. (2007). *Ending aging*. New York: St. Martin's Press.
Delio, I. (2014). *From Teilhard to Omega*. New York: Orbis Books.
Dennett, D. C. (1992). *Consciousness explained*. Boston: Back Bay Books.
Dennett, D. C. (1996). *Kinds of minds*. New York: Basic Books.
Dennett, D. (2009). Darwin's "strange inversion of reasoning". *Proceedings of the National Academy of Sciences of the United States of America, 106*(Suppl 1), 10061–10065.
Donald, M. (1991). *Origins of the modern mind*. Cambridge, MA: Harvard University Press.
Duncker, K. (1945). On problem-solving. *Psychological Monographs, 58*(5), 1–113.
Emerson, R. W. (2008). *The complete works of Ralph Waldo Emerson: Society and solitude* (Vol. 7). Boston, MA: Belknap Press.
Engels, F. (1946). The part played by labour in the transition from ape to man. In C. Dutt (Ed.), *Dialectics of nature* (pp. 279–296). London: Lawrence & Wishart.
Engels, F. (2015). *Socialism: Utopian and Scientific*. (E. Aveling, Trans.). CreateSpace Independent Publishing Platform.
Engeström, Y. (1993). Developmental studies of work as a testbench of activity theory: The case of primary care medical practice. In S. Chaiklin & J. Lave (Eds.), *Understanding practice* (pp. 64–103). Cambridge: Cambridge University Press.
Ferrando, F. (2013). Posthumanism, transhumanism, antihumanism, metahumanism, and new materialisms: Differences and relations. *Existenz, 8*(2), 26–32.
FM-2030. (1989). *Are you a transhuman?* New York: Warner Books.
Frischmann, B., & Selinger, E. (2018). *Re-engineering humanity*. New York: Cambridge University Press.
Fuller, R. B. (1938/1971). *Nine chains to the moon*. New York: Anchor.
Fuller, S. (2008). *Dissent over Descent*. Cambridge: Icon Books.
Fuller, S. (2011). *Humanity 2.0*. Basingstoke: Palgrave Macmillan.
Fuller, S. (2012). Social epistemology: A quarter-century itinerary. *Social Epistemology, 26*(3–4), 267–283.
Fuller, S. (2013). *Preparing for life in humanity 2.0*. Basingstoke: Palgrave Pivot.
Fuller, S. (2014). Neuroscience, neurohistory, and the history of science: A tale of two brain images. *Isis: An International Review Devoted to the History of Science and Its Cultural Influences, 105*(1), 100–109.
Fuller, S. (2015). *Knowledge*. The philosophical quest in history. Abingdon and New York: Routledge.

Fuller, S. (2017). Humanity's lift-off into space: Prolegomena to a cosmic transhumanism. In R. Armstrong (Ed.), *Star ark: A living, self-sustaining starship* (pp. 383–393). Chichester: Springer Praxis.
Fuller, S. (2018). The brain as artificial intelligence: Prospecting the frontiers of neuroscience. *AI & SOCIETY*, 1–9.
Fuller, S. (2019). The metaphysical standing of the human: A future for the history of the human sciences. *History of the Human Sciences, 32*(1), 23–40.
Fuller, S., & Lipinska, V. (2014). *The proactionary imperative*. Basingstoke: Palgrave Macmillan.
Funkenstein, A. (1986). *Theology and the scientific imagination from the middle ages to the seventeenth century*. Princeton: Princeton University Press.
Galilei, G. (1632/2001). In S. Drake & S. J. Gould (Eds.), *Dialogue concerning the two chief world systems*. New York: Modern Library.
Gazzaley, A., & Rosen, L. D. (2016). *The distracted mind*. Cambridge, MA: MIT Press.
Glanvill, J. (1931). *The vanity of dogmatizing*. New York: Cambridge University Press.
Greenwood, J. (2015). *Becoming human*. Cambridge, MA: MIT Press.
Hacking, I. (1975). *Why does language matter to philosophy?* Cambridge: Cambridge University Press.
Hahn, H. (1933). *Logik, Mathematik und Naturerkennen*. Verlag Gerold.
Hansell, G. R., & Grassie, W. (Eds.). (2011). *H+/−: Transhumanism and its critics*. Philadelphia: Metanexus Institute.
Harari, Y. N. (2017). *Homo Deus*. New York: Harper.
Haraway, D. (1991). A cyborg manifesto: Science, technology, and socialist-feminism in the late twentieth century. In *Simians, cyborgs and women* (pp. 149–182). New York: Routledge.
Harrison, P. (1998). *The bible, protestantism, and the rise of natural science*. Cambridge, MA: Cambridge University Press.
Harrison, P. (2007). *The fall of man and the foundations of science*. Cambridge, MA: Cambridge University Press.
Harrison, J. (2010). *Enhancing evolution*. Princeton: Princeton University Press.
Harrison, P., & Wolyniak, J. (2015). The history of 'transhumanism'. *Notes and Queries, 62*(3), 465–467.
Heersmink, R. (2015). Dimensions of cognitive integration in embedded and extended systems. *Phenomenology and the Cognitive Sciences, 14*(3), 577–598.
Heersmink, R. (2017). Extended mind and cognitive enhancement: Moral aspects of cognitive artifacts. *Phenomenology and the Cognitive Sciences, 16*(1), 17–32.
Hirsch, A., & Catchim, T. (Eds.). (2012). Come back, Peter; come Back, Paul: The relation between nuance and impact. In *The permanent revolution* (pp. 119–136). San Francisco: Jossey-Bass.
Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction, Special Issue on Human-Computer Interaction in the New Millennium, Part 2, 7*, 174–196.
HPlus Pedia (2019) Main Page. [Online] Available at: https://hpluspedia.org/wiki/Main_Page [Accessed 29 June 2019]
Hughes, J. (2004). *Citizen cyborg*. Cambridge, MA: Basic Books.
Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.
Hutchins, E. (2011). Enculturating the supersized mind. *Philosophical Studies, 152*(3), 437–446.
Huxley, J. (1957). *New bottles for new wine*. London: Harper & Brothers.
Iriki, A., & Taoka, M. (2012). Triadic (ecological, neural, cognitive) niche construction: A scenario of human brain evolution extrapolating tool use and language from the control of reaching actions. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 367*(1585), 10–23.
Kantorowicz, E. (1957). *The King's two bodies*. Princeton: Princeton University Press.
Kirsh, D. (2006). Distributed cognition: A methodological note. *Pragmatics & Cognition, 14*, 249–262.
Kirsh. (2009). Problem solving and situated cognition. In M. Aydede & P. Robbins (Eds.), *The Cambridge handbook of situated cognition* (pp. 264–306). New York: Cambridge University Press.

Kirsh, D. (2013). Embodied cognition and the magical future of interaction design. *ACM Transactions on Human-Computer Interaction, 20*(1), 30.
Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science, 18*(4), 513–549.
Kozulin, A. (1984). *Psychology in utopia*. Cambridge, MA: MIT Press.
Krueger, J., & Szanto, T. (2016). Extended emotions. *Philosophy Compass, 11*(12), 863–878.
Kurzweil, R. (1999). *The age of spiritual machines*. New York: Random House.
Kurzweil, R. (2005). *The singularity is near*. New York: Viking.
Logan, R. K. (2010). *Understanding new media*. Frankfurt/Main, New York: Peter Lang.
Lovejoy, A. O. (1936). *The great chain of being*. Cambridge, MA: Harvard University Press.
MacFarland, J. (2020). *The techno-centred imagination: A multi-sited ethnographic study of technological human enhancement advocacy*. Coventry: Palgrave Macmillan.
Malafouris, L. (2013). *How things shape the mind*. Cambridge, MA: MIT Press.
Malapi-Nelson, A. (2016). Transhumanism, Christianity and modern science: Some clarifying points regarding Shiffman's criticism of Fuller. *Social Epistemology Review and Reply Collective, 5*(2), 1–5.
Malapi-Nelson, A. (2017). *The nature of the machine and the collapse of cybernetics*. New York: Palgrave Macmillan.
Marmodoro, A. (2011). The metaphysics of the extended mind in ontological entanglements. In A. Marmodoro & J. Hill (Eds.), *The metaphysics of the incarnation* (pp. 205–227). Oxford: Oxford University Press.
Marx, K. (1906). *Capital: A critique of political economy*. New York: Random House.
Mazlish, B. (1993). *The fourth discontinuity*. Yale University Press.
McLuhan, M. (1964). *Understanding media*. Cambridge, MA: MIT Press.
Menary, R. (2007). *Cognitive integration*. Houndmills: Palgrave Macmillan.
Menary, R. (Ed.). (2010). *The extended mind*. Cambridge, MA: MIT Press.
Mercer, C., & Trothen, T. J. (Eds.). (2014). *Religion and transhumanism*. Santa Barbara: Praeger.
Mirandola, P. D. (1468/2016). In D. F. Borghesi, D. M. Papio, & D. M. Riva (Eds.), *Oration on the dignity of man* (Translation edition). New York: Cambridge University Press.
Moore, G. E. (1965). Cramming more components onto integrated circuits. *IEEE Solid-State Circuits Society Newsletter, 11*(3), 33–35.
More, M. (1996). *Transhumanism: Towards a futurist philosophy*. Retrieved from https://web.archive.org/web/20051029125153/http://www.maxmore.com/transhum.htm
More, M., & Vita-More, N. (Eds.). (2013). *The transhumanist reader*. Chichester: Wiley-Blackwell.
Nagel, T. (1986). *The view from nowhere*. New York: Oxford University Press.
Nayar, P. K. (2014). *Posthumanism*. Cambridge: Polity Press.
Newell, A., & Simon, H. A. (1976). *Computer science as empirical inquiry: Symbols and search. communications of the ACM, 19*(3), 113–126.
Noble, D. F. (1997). *The religion of technology*. New York: Penguin Books.
Norman, D. (1991). Cognitive artifacts. In J. M. Carroll (Ed.), *Designing interaction* (pp. 17–38). Cambridge, MA: Cambridge University Press.
Olson, C. E., & Meconi, D. V. (2016). Introduction. In C. E. Olson & D. V. Meconi (Eds.), *Called to be the children of God*. San Francisco: Ignatius Press.
Passmore, J. A. (1970). *The perfectibility of man*. London: Duckworth.
Persinger, M. A. (1991). Preadolescent religious experience enhances temporal lobe signs in normal young adults. *Perceptual and Motor Skills, 72*(2), 453–454.
Putnam, H. (1960). Minds and machines. In S. Hook (Ed.), *Dimensions of mind* (pp. 148–180). New York: New York University Press.
Putnam, H. (1967). Psychological predicates. In W. Capitan & D. Merrill (Eds.), *Art, mind and religion* (pp. 37–48). Pittsburgh: Pittsburgh University Press.
Ranisch, R., & Sorgner, S. L. (Eds.). (2014). *Post- and transhumanism*. Frankfurt/Main: Peter Lang.
Remedios, F. (2003). *An introduction to Steve Fuller's epistemology*. Lanham: Lexington Books.
Remedios, F. X., & Dusek, V. (2018). *Knowing humanity in the social world: The path of Steve Fuller's social epistemology*. London: Palgrave Macmillan.

Robinson, D. (2013). *Feeling extended*. Cambridge, MA: MIT Press.
Rowlands, M. (1999). *The body in mind: Understanding cognitive processes*. Cambridge, MA: Cambridge University Press.
Rowlands, M. (2010). *The new science of the mind*. Cambridge, MA: MIT Press.
Ruck, C., et al. (1979). Entheogens. *Journal of Psychodelic Drugs, 11*(1–2), 145–146.
Sandberg, A. (2013). Morphological freedom – Why we not just want it, but need it. In M. More & N. Vita-More (Eds.), *The transhumanist reader* (pp. 56–64). Chichester: Wiley-Blackwell.
Sandstrom, G. (2014). *Human extension*. Houndmills: Palgrave Pivot.
Savulescu, J., & Bostrom, N. (Eds.). (2009). *Human enhancement*. Oxford: Oxford University Press.
Shew, A. (2017). *Animal constructions and technological knowledge*. Lanham: Lexington Books.
Shiffman, M. (2015). Humanity 4.5. *First Things, 257*, 23–30.
Simon, H. A. (1981). *The sciences of the artificial (2nd Ed.)*. Cambridge, MA: MIT Press.
Slaby, J. (2016). Mind invasion: Situated affectivity and the corporate life hack. *Frontiers in Psychology 7*, 266.
Smart, P. R., Clowes, R. W., & Heersmink, R. (2017). Minds online: The Interface between web science, cognitive science and the philosophy of mind. *Foundations and Trends in Web Science, 6*(1–2), 1–232.
Sorgner, S. (2016). Three transhumanist types of (post)human perfection. In J. Hurlbut & H. Tirosh-Samuelson (Eds.), *Perfecting human futures: Transhuman visions and technological imaginations* (pp. 141–157). Wiesbaden: Springer.
Stiegler, B. (1998). *Technics and time: The fault of Epimetheus*. Stanford: Stanford University Press.
Sutton, J. (2007). Batting, habit, and memory: The embodied mind and the nature of skill. *Sport in Society, 10*(5), 763–786.
Sutton, J. (2010). Exograms and interdisciplinarity: History, the extended mind, and the civilizing process. In R. Menary (Ed.), *The extended mind* (pp. 189–225). Cambridge, MA: MIT Press.
Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data, 1*(2), 85–99.
Theiner, G. (2011). *Res cogitans extensa*. Frankfurt/Main, New York: Peter Lang.
Theiner, G., & Drain, C. (2017). What's the matter with cognition? A 'Vygotskian' perspective on material engagement theory. *Phenomenology and the Cognitive Sciences, 16*(5), 837–862.
Transhumanist FAQ (Version 3.0) [Online] Available at: https://humanityplus.org/philosophy/transhumanist-faq/ [Accessed 29 June 2019].
Vygotsky, L. S. (1978). In M. Cole et al. (Eds.), *The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
Vygotsky, L. S., & Luria, A. R. (1993). *Studies on the history of behavior*. Edited and transl. by V. Golod & Jane Knox. Hillsdale: Lawrence Erlbaum.
Weber, M. (1963). *The sociology of religion*. Boston: Beacon Press.
Webster, C. (1975). *The great instauration*. London: Duckworth.
Westfall, R. S. (2007). *Isaac Newton*. Oxford: Oxford University Press.
Williams, T. (2016). John Duns Scotus. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2016). Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/spr2016/entries/duns-scotus/
Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review, 9*(4), 625–636.
Wilson, R. A. (2004). *Boundaries of the mind*. Cambridge, MA: Cambridge University Press.
Wilson, R. A., & Foglia, L. (2017). Embodied cognition. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017). Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/spr2017/entries/embodied-cognition/
Wolf, G. (2009). Know thyself: Tracking every facet of life, from sleep to mood to pain, 24/7 365. *Wired Magazine*, 17.07. http://archive.wired.com/medtech/health/magazine/17-07/lbnp_knowthyself?

Georg Theiner is associate professor of philosophy at Villanova University (USA). He earned his PhD in philosophy, with a Joint PhD in cognitive science, from Indiana University. He has wide-ranging research interests in the philosophy of mind and cognitive science, social epistemology and social ontology, and the philosophy of language. He has published over 25 articles and book chapters in these areas and is author of the monograph *Res Cogitans Extensa: A Philosophical Defense of the Extended Mind Thesis* (Peter Lang, 2011). He is particularly interested in conceptions of "*e*mbodied, *e*mbedded, *e*xtended, and *e*nactive" ("4e") cognition, socially distributed cognition and the nature of epistemic collaboration, and more recently questions about artificial intelligence and its impact on the future of humanity. He currently serves as executive editor of the journal *Social Epistemology* (Taylor & Francis) and has refereed for over 25 professional journals from a variety of academic fields. He is passionate about exploring the use of innovative technologies in higher education.

Index

R. W. Clowes et al. (eds.), *The Mind-Technology Problem*, Studies in Brain and Mind 18, https://doi.org/10.1007/978-3-030-72644-7