This is a preprint. Please do not cite this version. Forthcoming.
Clowes, R., and Gaertner, K., **Hipólito, I.** (2021/in press). **The Mind-Technology problem and the deep history of mind design.** In Clowes, R., and Gaertner, K., **Hipólito, I.** (eds.) *The Mind-Technology Problem - Investigating Minds, Selves and 21st Century Artifacts.* Studies in Brain and Mind Springer International Publishing.

The Mind Technology Problem and the Deep History of Mind Design

Robert W. Clowes, Klaus Gärtner and Inês Hipólito

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 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 17th 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 17th 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 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

[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 putative immaterial substances might interact with the rest of the causal order. For us, it is a clear example of how it is possible to think against the grain of even a highly dominant conceptual scheme.

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 & Shields, 2019) 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 17th 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 17th 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

[2] Indeed, it is only since the 1960s that there has been – at least in the Anglo 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 20th 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.

problem have been resolved. Far from it. But at least for those working 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 seventy 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*

[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 & Turner, 2002; Lakoff & Johnson, 2003 [1980]), perhaps especially in (Lakoff & 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.

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 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 proposal 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 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.

[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)

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 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, this time, between the workings of human minds and the workings of the machines we create. The age of the fourth discontinuity is one

[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 speaks of the fourth revolution as the information revolution. Floridi does not mention Mazlish's (1993) formulation although there are interesting differences between the two we will discuss in the next section.
[10] An idea resisted by some (Gibson, 1979; Tallis, 2004; Varela, Thompson, & Rosch, 1991).

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.

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. 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 the possibility, the aspiration and crucially 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. Artificial Intelligence (AI) therefore unifies in one research program a wholesale reframing of what human minds are, at the same time as producing intelligent or "smart" systems that can apparently do autonomously much cognitive work.

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 when he writes: "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[12] may have been

[11] The early modern period may give us some of the most well-known examples of using our technologies as metaphors for our minds, such as the hydraulic metaphor used in the time of Descartes to illustrate the ways that bodies were supposed to work, to the Mills of Leibniz's thought experiments, to the 19th century idea that telegraph connections could form of a model of interconnection of brains.

[12] Expressed in Hobbes 1651 political treatise *Leviathan.* Luciano Floridi gives an interesting reconstruction of how Hobbes ideas may have been influenced (Floridi, 2014, p. 91)

triggered by Blaise Pascal's creation, in around 1642 - 44, of a calculating device – today known as the 'Pascalina' – capable of four arithmetical operations.[13]

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 of the workings of our minds in mechanist terms[14]. Boden argues that the notion is "more securely dated to the time of the second world war" (Boden, 2006, p. 52). Specifically, Alan Turing´s 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 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. AI then, first a theoretical possibility, and then a practical endeavor which is now reshaping the human world is what reconfigures the mind-body problem into the mind-technology problem.

The first stage of the mind-technology problem which is primarily the generation of a new set of theoretical problems for philosophers and scientists. It is bound up with artificial intelligence as 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 of our thought. 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

---

[13] Luciano Floridi gives an interesting reconstruction of how Hobbes ideas may have been triggered by the creation of this machine which was clearly influential throughout Europe.

[14] 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 2,500 years. 'Man as Machine' is much more recent." (Boden, 2006, p. 51).

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 good beat world champions in first Chess and then Go. IBMs 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. It arises as more of a practical challenge of how to live with our creation. Although AI applications have been increasingly permeating society over the last thirty or forty 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 in the developed world. As many of us now interact with "smart assistants" such as Amazon Alexa, a new form of existential challenge arises when the everyday world becomes populated with these creations[15].

The mind-technology problem can be regarded as the successor constellation to the mind-body problem which does not make the same default assumptions about mind. Whereas the mind-body problem tacitly assumes that the definition of mental processes is unproblematic and locates the basic difficulty in how our (ethereal) mental processes causally interact with matter, the mind-technology problem by contrast assumes that mental processes are material. Strictly speaking the problem interaction dissolves. 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 behind the mind-technology problem 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 give it 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?

[15] Hans Moravec (1988) calls them *Mind Children.* It is because this creation of at least would-be minds, and then our further co-habitation with our creations is a central problematic of our age that Mary Shelley´s (Shelley, 2018 [1818]) Frankenstein remains the prescient touch stone text of our epoch.

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 between the sorts of intelligent processes that are parts of minds and those that instantiated by the artefacts we create?

More practically the mind-technology problem grows out of the ability to model and build actual artefactual systems that can exhibit intelligent behaviors, that is, 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 artificially 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.

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, 1975; Newell & 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 Artificial Intelligence 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 Minskey, John Holland, Claude Shannon, Oliver Selfridge, Ross

Ashby, Allen Newell and Herbert Simon[16]. 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 early boldness of the vision; and indeed, the over-zealous timeframe for deliver was to proof something of 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). 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 allowed problems spaces to be-specified in symbolic ways. One limitation of this research – and arguably one that is still a major limitation today – was that despite the many impressive successes in building systems to reason about such pre-specified 'microworlds', 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).

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 respond by asking further plausible sounding canned questions, sometimes eliciting rather complex responses from its human interactants (Weizenbaum, 1976). 'Conversing' with Eliza was often reported as a valuable experience by some of its users

[16] 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 three day symposium convened at MIT in mid-September and a later meeting in held in 1958 in London as times and places where the central ideas of AI really germinated (see Boden, 2006, Chapter 6.iii).

who 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 features of human natural language.

It would be wrong to present all of this early work in AI as simulacra and smoke and mirrors, however. By the 1970s an extension of symbolic processing techniques, led to the *Expert System* paradigm which sort of symbolically encode and process expert knowledge and inference patterns in certain discrete domains. 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 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[17].

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 focused on the symbolic systems paradigm that was the core of much cognitive modelling in the early days. In addition, neural modelling – especially Rosenblatt's early work on the perceptron of algorithms – attempted to model intelligence at the level of the brain. This program was killed off for more than a decade by Minsky & Papert´s (1969) *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 & McClelland, 1986a, 1986b). From the early days in cognitive science, AI researchers were able to build many penetrating models of cognitive processes – including creative processes that were often held to be beyond the purview of mere machines (Boden, 1990) – that offered powerful models of human cognition. This work paved the way for today´s deep learning algorithms (Bengio,

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

2009) that supply much of the algorithmic bases behind Google´s search monopoly in so many areas of digital life. Brooks experiments in situated robotics (Brooks, 1991a, 1991b) provided the anti-thesis to the main thesis of computational approaches to AI. Today, rather than opposition, they seem to be part of a new embodied / predictive synthesis in the explanation of the mind (Clark, 2015a)[18].

In the midst of these early adventures in prospective Artificial Intelligence (AI), two intertwined and overlapping approaches or tendencies in research emerged that would shape much thinking – and launch a thousand controversies – about AI. 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.[19] 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, Breazeal, Marjanovic, Scassellati, & Williamson, 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 were only introduced somewhat later by John Searle in his attack on the idea that symbol process 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

[18] 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).

[19] 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 performing the laborious calculations that Pascal´s father needed to perform as part of his work as a supervisor of taxes.

faces of AI: the "weaker" engineering program, and the "stronger" scientific program continue to interlock up until the present day[20]. As we shall see however both 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 Section 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),[21] ideas of computers and computationalism provides much of the conceptual apparatus and motivation for functionalist thinking. Functionalism heavily lend 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[22]. 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

[20] Today the term Artificial General Intelligence (AGI) is sometimes used in order to describe artificial systems which have human level intelligence (e.g., Goertzel & 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 in Searle´s sense. It is conceptually possible to build an AGI that could match or even outperform human beings in any 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. Recent work in machine consciousness seeks, among other things, to attempt to understand if non-organic mechanistic systems – predominantly computational systems – could ever be subjective in Searle´s sense.

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

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

increasingly exotic set of realization bases, i.e. 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 Section 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), 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[23]. 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, 1975; Newell & 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, Thompson and Rosch, 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

[23] 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 lots of problems, not least when we consider the notion of personal identity (see the papers by Schneider and Corabi, and Piccinini, this volume).

and mental processes might be, and what kind of systems might be considered to instantiate them.

Thus, 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? 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) come to mind as puzzles that are at the core of the new understanding.

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. Not just Homo Faber, man the maker, but human beings the self makers[24].

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 21st 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 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, Cosmides, & Tooby, 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 of fashioning of artefacts,

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

is, once again, gaining ground (Ihde & 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). 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 then 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 & Malafouris, 2008; Vygotsky, 1962). Human beings, through the making of tools, open up new developmental trajectories 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 argued for 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?

What human minds are, therefore, strongly depends upon our nature as the users and creators of technology. 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, Odling-Smee, & Feldman, 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 cognitive abilities (Dennett, 1996b; Gregory, 1981). Moreover, according to recent theoretical developments, our artefacts are not just parts of our environments but can produce profiles that contrast with – according to the extended mind view – also part realize our mental states (Clark & Chalmers, 1998).

The Extended Mind Thesis (EMT) holds 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 under some conditions come to count as 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 in a way that promises a powerful 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 and fulfilled the famous trust and glue conditions could be considered a part of an individual´s mind then that system could be considered part of an individual´s mind[25]. 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 history and indeed 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 there are deep continuities in the history of our tool use and its implications for our cognitive development and the future of our minds.

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

Not everyone agrees with this perspective on technology. According to Floridi´s (2014) book *The Fourth Revolution*[26] 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 increasingly 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 adn 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 different. What are these changes? Our sense of self has indeed undergone a radical shift, if not so much the nature of human cognition and the mind. 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).

The concept of inforg is used to express a sort of parity to the ways human beings and the other information processing systems interact in the new terrain of "onlife", but importantly for Floridi, computers are just syntactic processors and lack the meaningfulness or original intentionality of human minds[27]. Our conceptions of mind are thus undergoing profound change even if the cognitive functioning 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.

The idea is that there is increasingly no 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; they are increasingly the natives of this realm. Floridi describes the way there is increasingly less

---

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

[27] In this, the foundations of Floridi´s view clearly echo John Searle´s position that no amount cokmputational syntactic processing can ever add up to the inherent meaningfulness and subjectivity of real minds. (Add citation to

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 lives experience (Floridi, 2015) [28]. The two apparently distinct realms are now so interdependent it is becoming increasingly difficult to separate them even in theory.

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 fifty 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 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, the cognitive adaptations which take place with each epoch with 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, (as just mentioned). If we take such cognitive fluidity to be a special feature of human cognition, and Mithen is right, one of the central attributes of the human mind arose in this 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

[28] 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.

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 4000BCE, 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 development 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, Petersson, Reis, Stone-Elander, & Ingvar, 1998; Dehaene, 2009).

In the late 20th and early 21st 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[29]. Floridi writes that hyperhistorical

[29] Albeit, see the discussion at the beginning of Section 3 on some of the cross-cultural diversity that there is in how human's have variously thought of AI.

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[30] 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 21st Century and difficult to imagine non-catastrophic circumstances where this is not also a dominant feature of any near-future human societies[31]. 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 20th Century and especially the early part of the 21st Century signal a new phase 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 another some aspects of AI technology into their design. The reasons are implicit

[30] 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.
31

in the foregoing discussion, namely that the technologies have long made a contribution to human intelligence in one form or another. However, there does appear to be several factors which are new and discontinuous with the history of technology. Central is that smart technologies do not merely contribute toward, or mediate human intelligence, but at least appears to be interpreted as 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?

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. Deep theoretical questions now move from the seminar room to be the practical challenges that confront human beings through 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 is: *How are we to live with smart artefacts.*

Forms of AI derived technology are becoming ever more pervasive in society from the invisible systems setting our credit ratings, helping us finding our way to a restaurant, or a date, to the conversational agens such as Amazon Alexa or Apple Siri that we consult on our mobile devices or the smart speakers that reside in our homes. At one level our interaction with AI in the form of smart technologies may seem to be essentially 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 on intelligences, whose ultimate status vis-à-vis our own, 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 should be regarded as type different from our own. He dubs smart systems as instances of Light AI (Floridi, 2014, p. 141) and that it is not really intelligent at all.[32] 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 more a sort of syntactic processing[33]. Our smart artefacts exhibit only weak (light) AI and are moreover not really intelligent. On the interpretation favored by Floridi, AI technology now and, in at least the near future, is likely to be Light AI. The technology is in a sense the inheritor of the Eliza program. While AI may accomplish useful tasks for us, it is primarily doing so by blind syntactic processes which, while they may be useful, share little with human cognitive processes. They are not really a form of intelligence at all. 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.

Regardless of its status, such Light AI 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 now 5G networks rapidly being deployed, as the internet of things is becoming an everyday reality (P. R. Smart, Madaan, & Hall, 2018), smart artefacts, however we interpret the nature of the intelligence they embody

[32] See Chapter 6, "Intelligence Inscribing the World" of (Floridi, 2014) for a detailed treatment of this theme and also 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).

[33] 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 doesn't help matters. Strong and Weak AI were originally used to distinguish two approaches what AI was supposed to be doing, either engineering or cognitive science. GOFAI and nouvelle AI were different approaches to how these goals could be obtained (See e.g., Brooks, 1990).

are increasingly ever-present parts of our lives. Even if these increasingly ubiquitous smart artefacts are not really independently intelligent, they may nevertheless be becoming parts of our intelligence, both collective, but also personal. 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. 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). This is pretty clearly a misrepresentation of 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 idea of the Extended Mind which may have seemed like a distance and exotic theoretical possibility when the idea was first mooted in 1998[34], 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[35].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[36]. 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

[34] See for instance the discussion in (Fodor, 2009).
[35] This is one reason the original extended mind paper is the most cited philosophy paper of the last 20 years.
[36] 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 either by replacement, extension or augmentation, see (Clowes, 2013) for further discussion and the slightly problematic nature of these definitions.

remember the actual information itself (Sparrow, Liu, & Wegner, 2011; Wegner & Ward, 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?[37] Our intimate encounters with and increasing reliance upon these technologies may thus be rapidly changing how we conceive of human memory (Clowes, 2017) and indeed other psychological involved human attributes such as personal identity (Heersmink, 2016). E-memory devices may provoke significant changes both in how we conceptualize memory, i.e., whether what we store on our phones can count as memory traces or not, but also has changed the structure of human memory, i.e. the various prompts etc. that we receive from the system (Clowes, 2013)[38].

In this way, the mind-technology problem becomes implicated in our folk-psychology, in how the folk conceives 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[39]. 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).

[37] 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.

[38] It is only in the time of cloud-tech 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.

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

What might be the consequences of these practices coming to accommodate our interactions with, for example, personal digital 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 as a sort of technological construction. 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 & 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 (P. R. Smart, Heersmink, & Clowes, 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, even if, or perhaps especially if, they should be understood as Light AI in the way Floridi suggests, can become part of our minds, we might view this as a less worrisome sort of interaction than the outsourcing picture where some of what was once our cognition is understood as becoming independent of ourselves. Which

picture, and under which circumstances, better fits the realities of dependencies on smart technologies is a nuanced and still much underexplored territory (Clowes, 2021).

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 (Clowes, In Press).

However, we come to resolve these problems, the boundaries between us and our machines, will be further breached in the coming years. Susan Schneider (Schneider, 2019) argues in her recent book that the technologies of cognitive augmentation are already much closer than we think. She raises the idea of a "Mind Design" salon where – in the near future – people will be able to enter to commission upgrades for their cognitive abilities. Schneider is particularly interested here in questions that we examine throughout this book, especially those 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.

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, is that 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 & Chiriatti, 2020a).

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.

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 really 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. A new wave of AI ethicists (Coeckelbergh, 2020; Taddeo & Floridi, 2018) argue that many of the scenarios entertained about AGI are closet science fiction. They distract us from clear thinking about the real ethical dilemmas we have with the much more limited but real AI of today. Yet, it remains generally unclear of what the status of AI is today, with some experts claiming we may be only one major innovation away from human level performance (Russell, 2019). The Mind-Technology problem is thus one that confronts human beings as an existential question even if it is perhaps best not to frame it merely as an existential risk.

If we have, up until now in this introduction, followed Freud in framing the challenges to the self-conception 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 major economically advanced nations have reflected on AI and how it effects 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 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 thirty years (Smart, Chu, O'Hara, Carr, & Hall, 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. It is far from clear that the co-existence of human beings with AI are universally treated with fear and loathing.

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 to 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[40]. 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. It is clear that China has already made and plans to make further significant investment in AI. The Chinese state has already unleashed massive spending following its State Plan in AI Factors including the SmileToPay, which uses face recognition algorithms to validate a citizen's identity through the *Social Credit Plan* (Smart et al., 2019). 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).

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. An important direction we have discussed here is that even weak AI, smart, rather than truly intelligent gadgetry can, if we come to deeply depend upon it, transform human cognition. 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 light 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, 2021), but also largely unconscious and haphazard experimentation is highly likely. Foreclosing 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

[40] For a nice description of how these developed see (Somers, 2019)

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)[41]. 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, while acknowledging that the technology is itself undergoing a rapid process of evolution and application to a wide range of problems increase. Having access to "light" AI that can, for instance, be tailored to write an article about AI for the Guardian newspaper changes indeed the way human beings are likely to apply and think about their own intelligence (Floridi & Chiriatti, 2020b)[42].

A third way between the various forms of alienation is to acknowledge the radical openness of the problem[43]. 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.

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

---

[41] 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, 2020).

[42] 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.

[43] See also the preface of this volume by Steve Fuller, (Fuller, 2011) and the concluding chapter from Georg Theiner for further reflection on this topic.

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, or indeed artefactual 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 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 & 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)[44]. 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

[44] 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[45]. This re-interpretation, 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 complex, especially if it follows the artificial general intelligence 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 would especially 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 embodied by, 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), Bratman (2007), Korsgaard (2009) and 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,

[45] See Chapter 7 of Dennett´s *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.

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.

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 to ask whether Intel-Mary is still phenomenally conscious and hence opens, at least, the door to ask the question of whether machines can be phenomenally conscious.

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*, argues that mind uploading is not a serious possibility. Piccinini notes that many people think that mind uploading should be feasible because it is often held that the mind is like a software program running on our brain (computer functionalism) (Piccinini, 2010). 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 in principle conceivable. 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 chapter 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 Schneider and Corabi, *Cyborg Divas and Hybrid Minds*, extend previous work on mind-uploading (Corabi & 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 & 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[46]. The paper begins with a discussion of the extended mind thesis (EMT) (Clark and Chalmers, 1998) and the authors's arguments 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 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 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 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. This is due to the idea that 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 "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 & Chalmers, 1998). The idea here is that a form of partial

[46] 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).

uploading may already – implicitly and largely unconsciously be taking place – 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, our deep and ongoing interactions with these systems. In this *slow continuous uploading* (SCU) scenario, Clowes and Gärtner ask, which aspects of personhood could persist through uploading? The paper, then, directly takes up the challenge from Schneider and Corabi (this volume) that such a partial upload could never turn out to be complete, since it would lack core cognitive and especially core conscious features. 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 a Dennettian (Dennett, 1991) or Clarkian (Clark, 2006) world, a clear core to consciousness and the self or person turns out to be more illusory than expected. To make their argument, the authors refer us to Smart's idea (this volume) of how such an "illusion of conscious self" could be implemented by a digital system. Finally, they admit that even SCU, as "psychological continuity", has to deal with discontinuities. However, Clowes and Gärtner argue, these discontinuities may not be more profound than the transitions in the life cycle of a butterfly that, despite very different incarnations throughout its life-stages, still counts as one and the same.

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 current machine learning approaches. 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. This means, assuming that the brain is really a prediction machine (Clark, 2015b; Hohwy, 2013), deep learning systems may be able to emulate the functions of the brain 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 giving us 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 & 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.

3.3 The Epistemology, Ethics and Deep History of the Mind Extension

The third section of the book discusses the extended mind thesis in the context of 21st century technology and the deep history of Western concepts of the mind. 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 these issues become extremely clear 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, Goodman, Reardon, & Thompson, 2014; Chappelle, McDonald, et al., 2014), 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 nature of 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 a second contribution, Lukas Schwengerer, in his chapter *Extending Introspection*, argues for the possibility of extended introspection. This idea is directly 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 introduces the classical extended mind scenario of Otto, which he later expands to the scenario of Otto++ (P. 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 21st 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, Clark, Kallestrup, Palermos, & Pritchard, 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[47]. 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[48]. 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 deep historical background of the Christian tradition. The paper sets the work of Andy Clark (Clark, 2003, 2008), and especially the theory of the extended mind, in the context of the history of the Christian view of human nature and embodiment. Theiner especially interprets the works of Steve Fuller (2011) on these themes. Theiner analyses, in detail, Fuller's 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

[47] See also Clark´s discussion of the principle of ecological assembly (Clark, 2008).
[48] See further discussion in (Clowes, 2015).

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 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.

4: The Mind Technology Problem 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. It´s 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 are where do our minds stop, and our artefacts begin and how do minds like ours emerge.

The second phase of the mind-technology problem has been underway now for at least ten 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 philosophical questions, but as a series of 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 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.
Barkow, J. H., Cosmides, L., & Tooby, J. (1992). *The Adapted Mind : Evolutionary Psychology and the Generation of Culture*: OUP.
Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and trends® in Machine Learning, 2*(1), 1-127.
Bickhard, M. H. (2009). The interactivist model. *Synthese, 166*(3), 547-591.
Boden, M. A. (1977). *Artificial Intelligence and Natural Man*.
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., & Sandberg, A. (2008). Whole brain emulation: a roadmap. *Lanc Univ Accessed January, 21*, 2015.
Bratman, M. (2007). *Structures of Agency: Essays*: Oxford University Press, USA.

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.
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 University Press.
Castells, M. (1996). The information age: Economy, society and culture (3 volumes). *Blackwell, Oxford, 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. (2014). 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. (2014). 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*. Camb. MA: MIT Press.
Clark, A. (2008). Supersizing the Mind: 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. Frankfurt am Main: MIND Group.
Clark, A. (2015b). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*: Oxford University Press.
Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis, 58*, 10-23.
Clowes, R. W. (2015). Thinking in the cloud: The Cognitive Incorporation of Cloud-Based Technology. *Philosophy and Technology, 28, Issue 2,*(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. doi:10.1007/s11097-018-9560-4
Clowes, R. W. (In Press). The Internet Extended Person: Exoself or Doppleganger? *Limité*.
Copenhaver, R., & Shields, C. (2019). 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.
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. (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*: Phoenix Books.
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*: 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*: OUP Oxford.
Floridi, L. (2015). *The Onlife Manifesto: Being Human in a Hyperconnected Era*: Springer Cham Heidelberg New York Dordrecht London.
Fodor, J. (1975). *The Language of Thought*. Cambridge, MA: MIT 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 Comput Biol, 4*(11), e1000211.
Fuller, S. (2011). Humanity 2.0: Foundations for 21st Century Social Thought: Palgrave Macmillan London.
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.
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.
Hohwy, J. (2013). *The predictive mind*: Oxford University Press.
Hutto, D. D. (2008). *Folk psychological narratives: The sociocultural basis of understanding reasons*: The MIT Press.
Ihde, D., & Malafouris, L. (2019). Homo faber revisited: Postphenomenology and material engagement theory. *Philosophy & Technology, 32*(2), 195-214.
Jackson, F. (1982). Epiphenomenal Qualia. *Philosophical Quarterly, 32*, 127-136.
Kahneman, D. (2011). *Thinking, fast and slow*: Macmillan.
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.
Korsgaard, C. M. (2009). *Self-constitution: Agency, identity, and integrity*: OUP Oxford.
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, England: 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, United Kingdom: 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, U.S.A: 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, England: Bradford Book, MIT Press.
Menary, R. (2014). Neural Plasticity, Neuronal Recycling and Niche Construction. *Mind & Language, 29*(3), 286-303.
Minsky, M., & Papert, S. (1969). Perceptrons: An essay in computational geometry. *MIT Press.*
Mithen, S. (1996). *The Prehistory of the Mind*. London: Thames Hudson.
Moravec, H. (1988). *Mind Children: The future of robot and human intelligence*. Cambridge, Mass: Harvard University Press.
Newell, A., & Simon, H. A. (1972). *Human problem solving*: Englewood Cliffs, NJ: Prentice-Hall.
Ney, A., & Albert, D. Z. (2013). *The wave function: Essays on the metaphysics of quantum mechanics*: Oxford University Press.
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. Kegan Paul, Trench, Trubner & Co: London.
Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson.
Sapolsky, R. M. (1997). *Junk food monkeys and other essays on the biology of the human predicament*: 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, NJ: Wiley- Blackwell.
Schneider, S. (2019). *Artificial You: AI and the Future of Your Mind*: Princeton University Press.
Schwitzgebel, E. (2019). Introspection. In E. N. Zalta (Ed.), (Winter 2019 Edition ed., Vol. The Stanford Encyclopedia of Philosophy).
Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417-457.
Shelley, M. W. (2018). *Frankenstein: The 1818 Text*: Penguin.
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: OUP.

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 Edition* (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.
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.
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, England: Bradford Book, MIT Press.
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. (1950). Computing Machinery and Intelligence. *Mind, 49*, 433-460.
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. M. (1950). 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 Mass: 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.
Armstrong, D. M. (1980). The causal theory of the mind.
Baggini, J. (2018). *How the world thinks: a global history of philosophy*: Granta Books.
Benzon, W. L. (2020). GPT-3: Waterloo or Rubicon? Here be Dragons. *Here be Dragons (August 5, 2020)*.
Boden, M. A. (1977). *Artificial Intelligence and Natural Man*.
Boden, M. A. (2006). *Mind As Machine: A History of Cognitive Science Two-Volume Set*. Oxford: Oxford University Press.
Bostrom, N. (2014). *Superintelligence*: Dunod.
Brooks, R. (1990). Elephants Don't Play Chess. *Robotics and Autonomous Systems, 6*, 3-15h.
Brooks, R. (2002). *Robot: The Future of Flesh And Machines*. Cambridge, Massachusetts: 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.

Chalmers, D. (1995). Facing up to the Problem of Consciousness. *Journal of Consciousness Studies, 2*(3), 200-219.
Clark, A. (2006). Soft selves and ecological control. In D. Spurrett, D. Ross, H. Kincaid, & L. Stephens (Eds.), *Distributed Cognition and the Will*. Camb. MA: MIT Press.
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, Issue 2,*(2), 261-296.
Clowes, R. W. (2019). Immaterial engagement: human agency and the cognitive ecology of the internet. *Phenomenology and the Cognitive Sciences, 18*(1), 259–279. doi:10.1007/s11097-018-9560-4
Clowes, R. W. (2020). 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, UK.: Routledge.
Clowes, R. W. (2021). The Internet Extended Person: Exoself or Doppleganger? *Limité. Límite. Revista Interdisciplinaria de Filosofía y Psicología*.
Coeckelbergh, M. (2020). *AI Ethics*: MIT Press.
Copenhaver, R., & Shields, C. (2019). History of the Philosophy of Mind, Six Volumes.
Dennett, D. C. (1978). Artificial Intelligence as philosophy and psychology *Brainstorms*. Montgometry, VT: 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.
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.
Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality*: OUP Oxford.
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. A. (1975). *The Language of Thought*. New York.: Harvard University Press.
Fuller, S. (2011). Humanity 2.0: Foundations for 21st Century Social Thought: Palgrave Macmillan London.
Gardner, H. (1985). *The Mind's New Science*. New York: Basic Books.
Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*: Houghton Mifflin.
Goertzel, B., & Pennachin, C. (2007). *Artificial general intelligence* (Vol. 2): Springer.
Hendler, J. (2008). Avoiding another AI winter. *IEEE Intelligent Systems*(2), 2-4.
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. doi:10.1007/s13347-018-0321-7
Ito, J. (2018). Why Westerners Fear Robots and the Japanese Do Not: Wired.
Jackson, F. (1982). Epiphenomenal Qualia. *Philosophical Quarterly, 32*, 127-136.
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--it-could-boost-growth-despite-population-decline-abe-says.html
Kim, J. (2006). *Philosophy of mind* (2nd ed.). Cambridge, MA: Westview.
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.

Laland, K. N., Odling-Smee, J., & Feldman, M. W. (2000). Niche construction, biological evolution, and cultural change. *Behavioral and Brain Sciences, 23*, 131-175.
Malafouris, L. (2013). *How Things Shape the Mind: A Theory of Material Engagement*. Cambridge, MA, U.S.A: MIT Press.
Mazlish, B. (1993). *The fourth discontinuity: the co-evolution of humans and machines*: Yale University 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.
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.
Newell, A., & Simon, H. A. (1972). *Human problem solving*: Englewood Cliffs, NJ: Prentice-Hall.
Piaget, J. (1954). *The construction of reality in the child*. New York: Basic.
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.
Schneider, S. (2011). *The Language of Thought: A New Philosophical Direction*: Mit Press.
Searle, J. R. (1980). Mind, Brains and Programs. *Behavioral and Brain Sciences, 3*(3), 417-457.
Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., . . . Bolton, A. (2017). Mastering the game of go without human knowledge. *Nature, 550*(7676), 354-359.
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.
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, England: 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.
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. M. (1950a). Computing machinery and intelligence. *Mind, 59*(236), 433-460.
Turing, A. M. (1950b). Computing Machinery and Intelligence. *Mind, 49*, 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.
Vygotsky, L. S. (1962). *Thought and Language* (E. Hanfmann & G. Vakar, Trans.). Cambridge, MA: MIT 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.
Weizenbaum, J. (1976). Computer power and human reason: From judgment to calculation.
Wootton, D. (2015). *The invention of science: a new history of the scientific revolution*: Penguin UK.

This is a preprint. Please do not cite this version. Forthcoming.
Clowes, R., and Gaertner, K., **Hipólito, I.** (2021/in press). **The Mind-Technology problem and the deep history of mind design.** In Clowes, R., and Gaertner, K., **Hipólito, I.** (eds.) *The Mind-Technology Problem - Investigating Minds, Selves and 21st Century Artifacts.* Studies in Brain and Mind Springer International Publishing.