2019

Thinking with things : An embodied enactive account of mind–technology interaction

Anco Peeters
University of Wollongong

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Thinking with things: An embodied enactive account of mind–technology interaction

Anco Peeters

This thesis is presented as required for the conferral of the degree:

Doctor of Philosophy

Supervisors:
Dr Patrick McGivern
Prof. Robert A. Wilson (University of Western Australia)

Examiners:
Prof. Ezequiel Di Paolo (University of the Basque Country)
Prof. Michael Wheeler (University of Stirling)

The University of Wollongong
School of Humanities and Social Inquiry

December, 2019

Declaration

I, *Anne Coenrard Pieter (Anco) Peeters*, declare that this dissertation, submitted in fulfilment of the requirements for the conferral of the degree *Doctor of Philosophy* from the University of Wollongong, is wholly my own work unless otherwise referenced or acknowledged. This document has not been submitted for qualifications at any other academic institution.

Some of the chapters included in this dissertation are reproductions of co-authored articles of which I am the first author. These chapters are preceded by a signed certification that details the contribution of each author.

Anco Peeters

December 9, 2019

"Do these authors despise no one? The book is remarkable for the open-mindedness and generosity of its interpretations; the authors have clearly paid as much good attention to those they are criticizing as to their favorites."

Daniel Dennett (1993, p. 124) reviewing
Varela, Thompson and Rosch (1991)

"He is surely the kind of philosopher I like to hang out with, but more important than that, he is the kind of philosopher who is likely to make a difference in the field."

Francisco Varela (1993, p. 126) reviewing
Dennett (1991)

Abstract

Technological artefacts have, in recent years, invited increasingly intimate ways of interaction. But surprisingly little attention has been devoted to how such interactions, like with wearable devices or household robots, shape our minds, cognitive capacities, and moral character. In this thesis, I develop an embodied, enactive account of mind–technology interaction that takes the reciprocal influence of artefacts on minds seriously. First, I examine how recent developments in philosophy of technology can inform the phenomenology of mind–technology interaction as seen through an enactivist lens. Second, I show how an enactive account of remembering can improve operationalizations of the memory palace mnemonic through virtual reality devices. Third, I draw on virtue ethics to argue that an enactivist approach allows us to better grasp the morally shaping aspects of artefacts by looking at social robots. Fourth, I fend off an underlying metaphysical concern about enactivism by arguing that an embodied, enactive account is compatible with the multiple realization of cognitive processes. This principle is often seen as a crucial test favouring accounts such as extended functionalism over enactivism and I argue that some forms of enactivism pass this test as well. Finally, I conclude by considering what the future relationship between enactivism and functionalism may have in store for the study of mind–technology interaction.

Samenvatting (Abstract in Dutch)

Technologische artefacten hebben ons de afgelopen jaren tot steeds intiemere manieren van interactie verleid. Toch is er verrassend weinig aandacht geschonken aan hoe zulke interacties, zoals met draagbare apparaten en thuisrobots, onze geest, cognitive capaciteiten en ons moreel karakter vormen. In dit proefschrift ontwikkel ik een belichaamde, enactieve benadering van geest–technologieïnteractie die de wederkerige invloed van artefacten op de geest serieus neemt. Ten eerste onderzoek ik hoe recente ontwikkelingen in de techniekfilosofie de fenomenologie van geest–technologieïnteractie, bekeken vanuit een enactief perspectief, kunnen informeren. Ten tweede toon ik aan hoe een enactief begrip van herinneren het operationaliseren van de geheugenpaleistechniek door middel van *virtual reality*-apparaten kan verbeteren. Ten derde betoog ik, op basis van deugdethische overwegingen in onze omgang met sociale robots, dat een enactieve benadering ons beter in staat stelt de moreel vormende aspecten van artefacten te begrijpen. Ten vierde weerleg ik een onderliggend metafysisch probleem voor enactivisme door te betogen dat een belichaamde enactieve benadering te verenigen is met de meervoudige realisatie van cognitive processen. Dit principe wordt doorgaans gezien als een belangrijke proef die voordeel biedt aan uitgebreide vormen van functionalisme ten opzichte van enactivisme. Ik betoog dat sommige vormen van enactivisme ook voor deze proef slagen. Ten slotte overweeg ik wat de toekomstige relatie tussen enactivisme en functionalisme kan betekenen voor het bestuderen van geest–technologieïnteractie.

Contents in brief

Contents

Preface

I must have built my first robot when I was about 14 or 15 years old. That sounds more impressive than it is. In that period, which must have been around 2001, I managed to convince my parents to finance a subscription on a Dutch robotics magazine. Each new issue came with a component to construct your own robot and, as such magazines are wont to do, the first one or two issues were freely distributed before subscribers had to pay a quite hefty fee. Yet I remained subscribed and managed to construct a cute-looking, blue-domed robot on wheels that happily followed the flashlight I used to illuminate its surroundings. I have been fascinated by technology for as long as I can remember, but that moment still stands out to me as a bit of a revelation: we are able to build things that move around of their own accord, and do so in a seemingly intelligent manner!

What motivates the present dissertation is my curiosity about not just technologies, but particularly about how technologies shape our existence as experiencing, moral, and sometimes even intelligent human beings. This can be illustrated with a relatively simple example. It is well-known that George R.R. Martin, author of the *A song of ice and fire* series that was famously televised as *Game of thrones*, writes his hulking tomes in WordStar 4.0 on an old DOS desktop computer without Internet connection. This ageing machine provides all the tools he needs and none of the distractions of the modern digital workplace. Martin the writer is deeply entwined with his tool of choice and would not have it any other way. Having written the present text, I can understand some of his concerns. Expanding on this theme, we find J.R.R. Tolkien, one of Martin's main sources of inspiration, relating his feelings about being temporarily deprived from using his right hand. In a letter to his publisher Stanley Unwin dated October 1963, Tolkien laments how he found "not being able to use a pen or pencil as defeating as the loss of her beak would be to a hen" (letter

#248 in *The letters of Tolkien*). Note how Tolkien specifically mentions not being able to use the hand *for writing*. As becomes clear from his other letters, writing was for Tolkien as essential as breathing. But why are their specific writing implements of such importance to Martin and Tolkien? Why can't Martin just use a modern computer and why didn't Tolkien rely on a typewriter or friend to write things down?

Philosophy is able to shed light on such questions. Both Richard Menary (2007b) and Don Ihde (1990, p. 141) discuss the ways in which specific writing implements not only shape the act of writing – some tools allow faster writing than others – but actually change the author's mental activity and, therefore, the text being written. Coming from a philosophy of mind perspective, Menary argues that writing is, quite literally, thinking. Putting words and sentences on a piece of paper allows us to manipulate them in ways we couldn't do without the paper, which in turn feeds back into the writing process. Simply put: I am able to write part of an argument down and, when I reach the conclusion, restructure some of the original parts once I am clear on the exact steps involved in the argument.

Coming from a philosophy of technology perspective, Ihde reflects on the differences between using an old-fashioned dip or fountain pen, a typewriter, and a modern computer. Each enable different writing speeds and incline the author to different styles of editing. A dip pen invites one to write slowly and leaves room for thinking more carefully about words while writing. A modern computer, on the other hand, allows one to write fast and edit at whim. The reflections by Menary and Ihde are confirmed in an empirical study which revealed, among other things, that people who write with a pen generally think at the level of sentences and paragraphs while writing, and edit after the text is done. In contrast, authors who use computers are prone to pause and edit at the individual word level (Van Waes & Schellens, 2003). In this light, the paradoxical feeling of only knowing what one wanted to write by the very act of writing it, becomes more clear. This helps explain why authors such as Martin and Tolkien are so attached to their specific writing implements: the writing activity itself would otherwise likely be very different, as would the texts produced.

Reflecting on different types of writing implements may seem like a fairly innocent exercise, but it reveals something very fundamental about the nature of our engagements with various technologies. After all, if

merely exchanging a pen for a computer results in a different cognitive process and a different text, then to what extent do other, more complex technologies influence our thoughts and behaviours? And if our mental and bodily activities may depend so heavily on the characteristics of specific technologies, then to what extent do they become part of who we are? Indeed, many of us make photos of important events in our lives and, after many years, might not even remember such events if they did not have access to their photos. This has ethical implications as well. What if a person passes away and their friend or relative is responsible for going through their belongings: is the act of throwing away pictures kept by the deceased an act of memory removal, or perhaps even erasure of that person?

These are some of the questions with which the present dissertation engages. I wish for it to shed some light on these matters and clarify, even if only in part, the different ways in which we engage with technologies. If we are better able to understand such relations, then we are in a better position to hold on to human freedom and responsibility in a world where technologies are not only becoming increasingly complex but also increasingly invisible.

List of Figures and Tables

Introduction

Cognitive science is changing. It is witnessing a turn towards pragmatic, action-oriented, and dynamic approaches to cognition and away from views leaning on representations, computations, and mechanisms (Engel, Maye, Kurthen & König, 2013; Menary, 2016). Dramatically, philosopher Andy Clark (2016) says approaches that place context and action centre stage require us "to abandon the last vestiges of the 'input–output' model" (p. 139). In the present work, I look at the pragmatic turn through the lens of mind–technology interaction.[1] This move secures a double treasure. First, thinking about how minds engage with technologies in light of the pragmatic turn, helps us reconsider current approaches to cognition. Second, applying cognitive accounts that are at the vanguard of the pragmatic turn to cases of mind–technology interaction, helps us to better understand such types of interaction. These rewards make the present work of interest both to those working in cognitive science and to those working in mind–technology interaction, broadly construed.

My investigation of mind–technology interaction targets a crucial assumption in some prominent theories that attempt to align themselves with the pragmatic turn. To draw out this assumption, let us look again at Clark's input–output model. This model permeates many of the disciplines which constitute cognitive science, such as philosophy, psychology, cognitive neuroscience and artificial intelligence. This is unsurprising, because the input–output model is fuelled by the idea that our mind works like a computer. In the words of Paul Thagard (2005), in his *Mind: Introduction*

[1] While terms like 'human–computer interaction', 'human–technology interaction', or 'human–robot interaction' are perhaps more familiar, I deliberately use mind–technology interaction. *Mind*–technology interaction, because my research concerns the cognitive aspects of how we engage with environmental resources, such as technologies. Mind–*technology* interaction, because I would like to emphasise the general category of technological artefacts and not just those artefacts that compute.

to cognitive science: "Many but not all cognitive scientists view thinking as a kind of computation and use computational metaphors to describe and explain how people solve problems and learn" (pp. 3–4).[2] The computer metaphor is the assumption I target in the current investigation. It is a powerful metaphor: amongst its virtues is the concrete and mathematically precise toolkit that scientists and philosophers have used to study mind and cognition. However, the adequacy of this metaphor is up for debate and questioning it has important consequences for how we think about mind and cognition.

The computer metaphor underlies much recent work in understanding mind–technology interaction. Its most influential incarnation is that of *functionalism*: the philosophical idea that mental states are defined by what they do and not by what they are made of. Clark's input–output model is one form of functionalism. In contrast, the pragmatic turn is exemplified by the various strands of *enactivism*: the idea that mind arises out of an organism's active and continuing engagement with its environment. These two schools of thought form the guiding frames within which I examine mind–technology interaction.

In order to perform this examination, I will do the following. First, I show that the philosophical theory of functionalism underlies many of the current debates on mind–technology interaction. Second, I provide reasons to think that functionalism is not always in the best position to explain such interactions. Third, I argue that the competing theory of enactivism is better equipped to help us understand the relation between mind and technology. In sum, the guiding question motivating the present dissertation is whether functionalism makes any explanatory contribution over enactivism within the field of mind–technology interaction and, potentially, beyond. To situate these terms and steps, I now turn to a famous example from the philosophical literature on the interaction of mind and artefact: the case of Otto's notebook.

[2] There is little reason to doubt the *communis opinio* has changed much since Thagard's declaration. A recent announcement for the 2017 meeting of the *Cognitive Science Society* reads: "computation can serve as the foundational theory of how people actively process information in service of control and decision making … greater effort must be made to connect cognitive science theories to computational foundations" (cited in Núñez et al., 2019, p. 7).

In their classic paper "The extended mind", Andy Clark and David Chalmers (1998) take a closer look at how the mind may make use of environmental resources. They do so by considering a thought experiment that features two rememberers: Inga and Otto. Inga and Otto are both looking to visit the Museum of Modern Art while in New York. But whereas Inga uses her biological memory to recall the museum's address, Otto, a sufferer of early-onset Alzheimer's, retrieves it through his notebook. Otto's notebook, Clark and Chalmers argue, plays for Otto the same role that memory plays for Inga (p. 13). In it, Otto stores the things he would like to remember and his daily routine depends systematically, not incidentally, on his writing – similar to how Inga depends on her biological memory. Since Otto's notebook is playing the same role as Inga's biological memory and because we consider Inga's biological memory to be part of her mind, saying that Otto's notebook is not part of Otto's mind would be a case of neural chauvinism. Thus we have the extended mind thesis: that, in principle, the physical underpinnings of the mind may sometimes extend beyond the skull and skin.

While an excellent paper, that generated many research programmes on extended cognition, the proposal it puts forward is not without flaws. One important aspect, which remains underdeveloped from a mind–technology interaction perspective, is the fact that there are different ways to interface with technological artefacts and that those differences fundamentally matter. This is reason for Helena De Preester (2011) to warn us against equating Otto and Inga all too easily, arguing that, because "Inga has a stronger ownership over the informational items in her memory than Otto, and ... therefore the information in her memory functions differently from the information Otto has in his notebook" (p. 134). The different levels of ownership are themselves based on the different phenomenologies that Inga and Otto presumably have of the ways they bodily engage with their respective memories.

The role of the body is crucial if we want to be in a position to understand interfacing with artefacts and De Preester grabs hold of the right thread in this conceptual knot. But she is pulling in the wrong direction. If we are to untangle the problem, we should not follow Clark and Chalmers' lingo of "informational items" being accessed from storage, whether biological or otherwise. I take a cue here from Tom Froese (2014) who suggests

that "current symbolic computer interfaces are a source of alienation because their underlying principles are inherently alien to those of embodied life and mind" (p. 555). Though not a computer interface, Otto's notebook is vulnerable to the same critique. While interacting with artefacts through symbolic representations is certainly one way of interfacing, it need not be the only one. More fundamentally and perhaps paradoxically, interaction through symbolic representations might not even be best explained in terms of a computational theory of mind which puts information pick-up and processing at the base of cognition, as such theories have troubles accounting for the different roles the body plays. This is what I aim to show in the rest of the present dissertation.

Why is the notion of information processing central to many of the debates on extended cognition? This is due, I think, to the close connections between the extended mind thesis and the philosophical theory of functionalism. Functionalism is the textbook framework for understanding minds in analytic philosophy and cognitive science (Churchland, 2005; Brook, 2009). Since its (re-)conception in the 1960s, it has developed into many and sometimes contradictory shapes so it is understandable that Thomas Polger (2004) reports that "[v]arieties of functionalism are as varied as fingerprints but not nearly so constant" (p. 71). Bearing this complex history in mind, we may say that to a first and rough approximation functionalism describes mental states not by what they are made of (e.g., neural states), but by what they *do*. In a traditional form of functionalism, mental states are cast as computations over input, like sensory representations, which lawfully and structurally result in output, like motor actions. In an oft-repeated credo: "the mind is to the brain, as software is to hardware" (for a recent iteration see Piccinini, 2010). This mental manipulation of representations is what gives rise to the idea of mind as an information processor (Harman, 1988; Wilson, 1994).

The bridge between the extended mind thesis and functionalism rests on what Clark and Chalmers have dubbed the Parity Principle. This principle can be seen as a rule-of-thumb to determine when the mind extends: "[i]f, as we confront some task, a part of the world functions as a process which, were it done in the head, we would have no hesitation in recognising as part of the cognitive process, then that part of the world is (so we claim) part of the cognitive process" (Clark & Chalmers, 1998, p. 8).

Now, recall that functionalism explains mental states as requiring law-like, causal relations between systemic inputs and outputs. However, there is no need to understand these relations as necessarily being instantiated inside skull and skin. Indeed, functionalism does not even assume physicalism: witness Hilary Putnam (1967), one of functionalism's initial architects, exclaim how the doctrine "is *not* incompatible with dualism!" (p. 436). This theoretical flexibility leads Michael Wheeler (2015) to say that "functionalism plausibly provides a theoretical backdrop for the operation of the parity principle" (p. 160). It seems that the possibility of minds extending is a built-in feature of functionalism.

We are now in a better position to see just why it is "a common move in the literature to link [extended mind] in some way to ... functionalism" (Wheeler, 2015, p. 160). Both Clark and Wheeler have, in a formal bond between the two ideas, advocated a position Clark christened *extended functionalism* (Clark, 2008a; Wheeler, 2010a, 2010b, 2017). With the advent of the extended mind thesis in discussions on mind–technology interaction, its functionalist credentials have entered those debates as well (Aydin, 2012, 2015). Whether or not these functionalist commitments best serve discussions on mind–technology interaction in cognitive science is, as I argued earlier, a live question and, given the staying power of the extended mind thesis (Gallagher, 2018), an important one. To assess the viability of functionalism in debates on mind and technology, it is useful to compare it with a rival theory of mind.

A few years before the extended mind paper appeared, Francisco Varela, Evan Thompson and Eleanor Rosch (1991) presented the enactive approach to mind and cognition in their book *The embodied mind*. Inspired by work on self-producing or 'autopoietic' systems in molecular biology, the phenomenological tradition, and Buddhist ideas on the mutual dependence of cognition and the experienced world, enactivism understands mind as arising from an organism's active participation with the environment it's coupled with. In contrast with more traditional functionalist programmes, Thompson (2007) proposes that the nervous system is to be understood as an autonomous dynamic system which creates its own coherent patterns: the "nervous system does not process information in the computationalist sense, but creates meaning" (p. 13). Furthermore, where extended functionalism allows for minds to extend sometimes, enactiv-

ism holds that basic forms of cognition always have such reach (Hutto, Kirchhoff & Myin, 2014).

The emphasis on cognition as depending on active bodily engagement with an organism's environment and its inherently extensive nature puts enactivism in a better starting position than functionalism to explain the role of the body in mind–technology interaction. However, enactivism faces its own challenge as pointed out by Ezequiel Di Paolo (2009) when he admits that the "more interesting and forward-looking themes introduced by the [extended mind] approach, and towards which enactivism must still develop, include the problem space of technical individuation and technological networks that bootstrap the generation of cognitive identities" (p. 20). It is my hope that the present dissertation will make contributions to this development by providing an enactivist alternative to the extended functionalist tale of how minds and artefacts interact.

Some caveats have to be made regarding my depiction of functionalism up to this point. First, I must emphasise that Clark is, and has been, a staunch advocate of the importance of the body in cognitive science and, in his own words, would be horrified "to find myself suspected ... of now believing that the body didn't matter and the mind was something ethereal and distinct" (Clark, 2003, p. 189). Yet, good intentions notwithstanding, this might be exactly what he is doing. With the provocative accusation of 'body snatching', Shaun Gallagher (2015) draws attention to a recent trend by cognitive scientists who, in an attempt to march alongside the banner of embodied cognition, have infiltrated research programmes on embodied cognition and relegated any relevance the body might have to bodily representations *inside the brain*. Gallagher does not mention Clark in this colourful analogy, but Lawrence Shapiro (2019) does when he criticises Clark for focusing too much on the body as a computational resource instead of as a shaper of cognition. Regardless, the treatment of the body in extended functionalism deserves closer scrutiny and will receive it in Chapter 2.

The second caveat on my discussion of extended functionalism is that I have so far neglected to mention the evolution of the extended mind thesis through different 'waves'. The defining feature of the first wave of extended cognition was the Parity Principle, and in the second wave this principle was joined by the Complementarity Principle. Where the

former focused on similar cognitive functions being performed at different locations, the latter creates space for the idea that "different components of the overall (enduring or temporary) system can play quite different roles and have different properties while coupling in collective and complementary contributions to flexible thinking and acting" (Sutton, 2010, p. 194). These two principles are not mutually exclusive. The second wave signifies a change towards a more distributed type of thinking about cognition. It seems that the third wave, currently still only anticipated (Kirchhoff, 2012), will take this even further, with Gallagher (2018) suggesting that this wave may strive to integrate ideas about the brain as a predictive engine – the so-called 'predictive processing' paradigm – and enactivism into one coherent framework.

It is with the suggestion of integrating enactivism and extended mind that we have stumbled upon a deep question about the relationship between functionalism and enactivism. If integrating enactivism and extended mind is a live option, and, as we have seen earlier, extended mind and functionalism are seen as star-crossed, does that imply that functionalism and enactivism are on some level compatible with each other? Will the diametrical opposition I have presented between these two rivals collapse? In his introduction to the 2010a volume on the extended mind, Richard Menary seems to think such compatibility is an option: "It may turn out that a liberal functionalist account of cognition will provide a way of determining which manipulations are part of cognition and which are not, in which case there may not be any great tension between the enactive and functionalist approaches to the extended mind" (pp. 21–22). But he concludes by saying that the details of such a conjunction are not yet explicated.

Interestingly, proponents of both functionalism (Wheeler, 2010a, 2017) and enactivism (Di Paolo, 2009; Thompson & Stapleton, 2009) have denied that compatibility between the two approaches is possible – though they have done so for different reasons. Wheeler (2017) has gone so far as to say that extended functionalism is explanatorily superior to enactivism, particularly the branch of enactivism known as extensive enactivism. However, in recent work, Gualtiero Piccinini (2008, 2010, 2015) has taken steps towards disentangling functionalism from its traditional commitments to representational, computational, or mechanical theories of cognition. This

raises the question whether a pure version of functionalism is compatible with enactivist theories. One possibility is that enactivist theories in fact entail such a pure functionalism. If so, the functionalist framework must feature, despite appearances, in any such pragmatic approaches to cognition. This would mean that the denial of compatibility is based on a confusion of pure functionalism with a computational, representational, or mechanical account of cognition.

Even if some form of functionalism is entailed by enactivism, it leaves open the question of whether extended functionalism, if cast in the guise of a purified functionalist theory, makes an explanatory contribution to enactivism, *contra* Wheeler. Responding to this question will be the job of the remainder of this dissertation and I will do so by developing an enactivist account of mind–technology interaction and applying this account to specific instances of human–technology relations.

In Chapter 1, I engage with current discussions in the field of philosophy of technology. The *postphenomenological* approach is an important school of thought in contemporary philosophy of technology (Ihde, 1990; Verbeek, 2005; Rosenberger & Verbeek, 2015). Unlike classic phenomenology, postphenomenology does not see technological artefacts as an alienating factor between subject and world, but instead understands such artefacts as mediating between the two. For example, the discovery of ultrasonic images allows expecting parents to see the developing fetus. However, being able to see into the womb also confronts the parents with new ethical questions, such as when the fetus displays a painful chronic illness. Postphenomenology aims to clarify and structure the different force-fields – surrounding subjects, artefacts, and the wider world – that result from technological mediation.

There have been some attempts at converging extended cognition thinking and postphenomenology. Some have argued that both approaches are irreconcilable, as extended cognition assumes a subject–object dichotomy while postphenomenology does not (Kiran & Verbeek, 2010; Aydin, 2012, 2015). Others have attempted to show that extended cognition thinking is not vulnerable to such a critique (Heersmink, 2012). Enactivism similarly aims to understand mind and world as co-constitutive. Given that both enactivism and postphenomenology can count classic phenomenology amongst their pedigree, this will not be a surprise. My aim in Chapter 1,

therefore, is to show that enactivism is better placed to provide a cognitive framework to postphenomenology than extended functionalism. I do this by looking at the categorisation of different human–technology relations that have been put forward by postphenomenologists (Verbeek, 2008). This will provide enactivists with a robust theory of artefact engagement, while at the same time linking postphenomenology to an empirical cognitive theory.

Virtual reality has recently drawn the attention of some prominent philosophers of mind, as it may allow the investigation of scenarios which could previously only be imagined (Chalmers, 2017; Metzinger, 2018). One area where virtual reality opens up interesting lines of research is that of memory and mnemonics (Michaelian, 2016; Heersmink, 2018). In Chapter 2, I take a closer look at one of the most enduring and powerful mnemonics: the memory palace. Because mastering the memory palace takes a lot of commitment and practice, cognitive scientists have tried to support the technique through virtual reality, hoping to improve its accessibility. However, such operationalizations have so far not yielded results which can compete with traditional memory palace usage. I propose that current approaches to the virtual memory palace are based on an extended functionalist framework and, consequently, do not sufficiently account for the user's active bodily engagement in the memory palace. Instead, I develop an enactive account of the memory palace and recommend how future virtual operationalizations may benefit from design choices inspired by my enactive proposal. If my design recommendations are taken in by cognitive scientists and hold firm, we have further support for thinking enactively about memory in general.

Enactivism may have important ethical implications for how we think about ourselves. Chapters 3 and 4 form a pair that examines these implications within the context of social robotics. The first of these deals with the issue of sex robots. Though sex robots as such do not yet exist and are little more than sex dolls, manufacture of such devices is looming on the horizon. Unsurprisingly, major debates have erupted in society and academia about the use and implications of such robots. I propose that the framework of virtue ethics is well-disposed to examine the consequences of sex robots for human moral character. In doing so, I argue against current instrumentalist approaches to sex robot use. A contribution of the chapter

to the field of social robotics is that it suggests that, within supervised, therapeutic scenarios, it may be useful to implement robots with consent modules. This suggestion is not without risks, but the topic nonetheless deserves careful consideration before it is dismissed a priori. The chapter concludes with a reflection on the implications of sex robots for human autonomy and responsibility.

Because of the situated nature of virtue ethics, it is particularly of interest to enactive cognition researchers. Chapter 4 therefore immediately follows upon the issues raised in the previous and investigates the possibility of an enactive self and moral character. Situationists argue that humans can never be truly virtuous, saying that moral character is not consistent and overly dependent on environmental factors. I target their assumption that character and environment need be seen as strictly separate and defend the proposal that social robots can not only cultivate vice but also virtue. I do this by extending the concept of moral character to allow for the incorporation of environmental resources. I consider both extended functionalist and enactivist accounts of such an extended self, concluding that the latter provides a fundamentally more robust alternative. Thinking enactively about the self and moral character not only gives us ground for concluding social robots may support the cultivation of virtue, but also provides a novel answer to the situationist challenge to virtue ethics.

Finally, Chapter 5 examines a lingering issue between functionalism and enactivism, namely the principle of multiple realization. Because of enactivism's sensitivity to the concrete embodiment of cognitive acts, some are inclined to think that enactivism and multiple realizability do not play well together (Myin & Zahnoun, 2018). However, such considerations turn on an assumption that multiple realization is dependent on the conception of cognition as information-processing. I argue that there is an understanding of multiple realization that considers cognitive processes as potentially realized in cognitive systems, like humans in their own habitats, that systematically incorporate environmental resources. Contrary to what is often claimed, I argue that enactivism is compatible with multiple realizability and conclude that this principle thus need not give functionalism any decisive advantage over its competitor.

The internal logic of the present work is as follows. Broadly speaking the following chapters are connected as follows: Chapter 1 provides

a broad framework for thinking of mind-technology interaction as enactive. It connects, on a general level, the philosophy of mind and the philosophy of technology, and establishes the viability of an embodied approach to mind–technology interaction by drawing on phenomenologically inspired developments in both fields. Its major conclusions are that mind-technology interaction need not be understood in terms of informational exchange and that such interactions are co-shaping both agent and technology design.

This framework then informs discussions in Chapters 2, 3 and 4, where I apply the distinctions drawn in the first chapter to concrete case-studies. In Chapter 2, I focus on a specific mental process – namely memory – in relation to a specific technology – namely virtual reality. By showing how virtual interactions can be understood in embodied, enactive terms instead of information-processing ones, I not only provide proof of the pudding prepared in Chapter 1, but also aim to illustrate the ability of my framework to inspire technology design and cognitive science. Similarly, Chapters 3 and 4 together investigate the ethical implications of my framework. As I advance a theory of mind-technology interaction that takes the reciprocal influence of agent and environment seriously, some immediate points of ethical interest arise. For instance, if agent and environment are as intimately linked as I advocate in Chapter 1, moral responsibility cannot be thought to solely reside on the side of the agent. This calls for a ethical theory that is sensitive to the concrete and unique contexts in which moral acts take place. Virtue ethics is such a theory, as it reserves a prominent place for the way moral acts shape a person's character. Within the context of current debates on social robotics, the ethical implications of my framework are therefore made explicit in Chapters 3 and 4.

With the middle chapters securing the positive support for my argument, the fifth and final chapter aims to pre-empt a potential critique to it. The augmentation of postphenomenology by its alliance to the enactive approach puts it in a stronger position of relevance for cognitive science. With this, however, postphenomenology inherits a potential problem that has faced enactivists. Namely, if cognition is to be understood as embodied and enactive, as I claim in Chapter 1, it stands to lose the ability of information-processing cognitive approaches to understand cognitive processes as potentially implemented in different media. In Chapter 5, I

present this potential issue and show that my embodied, enactive framework can claim a similar ability to carve out the instantiation of a cognitive process in distinct materials

Current debates about wide cognition have been intense and recent years have seen a veritable explosion of literature on the topic. Inevitably, this means that the present work cannot investigate all possible avenues related to its main aim. Discussions about niche construction and scaffolding (Sterelny, 2010), cognitive integration (Menary, 2007a), cognitive archeology (Malafouris, 2013; Ransom, 2019), feminist theory (Ihde, 2002; Brancazio, 2019), and predictive processing (Hohwy, 2013; Clark, 2016) are all viable candidates for future dialogue partners with the present work. But concessions to the scope of the dissertation had to be made and I have attempted to restrain in-depth theoretical discussions to a critical pairing of extended functionalism and enactivism. I will leave it to future work to rectify any omissions this may have caused.

Though wide approaches to cognition are the talk of the day, not everyone has jumped on the train. Such thinkers that remain sceptical of cognition as extended or enactive argue that there is little to be gained from an explanation of mind as realized in part by environmental factors (Adams & Aizawa, 2001, 2008; Rupert, 2004). Wide cognition theorists have responded both with theoretical and empirical counterarguments (Menary, 2010b; Wagman & Chemero, 2014). This is an important discussion but as this dissertation is situated at the vanguard of discussions about wide cognition, it will not engage with these fundamental issues and instead assume that some form of wide cognition is a live possibility for theories of mind.

The methodology of the second part of this dissertation is likely somewhat different from what would commonly be expected in the context of analytic philosophy. This is deliberate. Inspiration for this methodology hails from philosophers who closely engage with empirical research and reach across the disciplinary boundaries within and outside of philosophy. Some of my philosophical heroes in this regard are Daniel Dennett and Andy Clark. My attention was recently drawn to a paper by Eric Schliesser (2019), in which he adopts the term 'synthetic philosophy'. His description is worth quoting in full:

> By 'synthetic philosophy' I mean a style of philosophy that brings together insights, knowledge, and arguments from the special sciences with the aim to offer a coherent account of complex systems and connect these to a wider culture or other philosophical projects (or both). Synthetic philosophy may, in turn, generate new research in the special sciences, a new science connected to the framework adopted in the synthetic philosophy, or new projects in philosophy. (pp. 1–2)

It is my intention that the present dissertation can be read in the spirit of synthetic philosophy as described in the quotation above.

The chapters in this work have been written as independent publishable papers and indeed some of them have appeared in academic journals. The publication status of each chapter is signalled at the start of the chapter where an abstract for that chapter can also be found in the style of an academic paper. While all chapters are, naturally, thematically connected and deal with related issues, some variation in tone and presentation has been unavoidable, particularly in those chapters that were co-authored. I have striven to keep the format of the chapters as consistent as possible. I trust this will not cause much inconvenience for the reader and apologise for any potential instances where it does.

My hope is that the present dissertation makes a convincing case for two points. First, that it shows how, in the domain of mind–technology interaction, an enactive approach to cognition helps inform and move forward some of the current discussions in that field. In particular, I have aimed to contribute to the design of new technological artefacts, such as virtual reality devices. Second, that it spurs on discussion between functionalists and enactivists about the future of their respective research programmes. Progress is often driven by opposition and I would like to see functionalists pick up the ball that I have kicked into their camp.

Abstract

In this paper, we evaluate the pragmatic turn towards embodied, enactive thinking in cognitive science, in the context of recent empirical research on the memory palace technique. The memory palace is a powerful method for remembering yet it faces two problems. First, cognitive scientists are currently unable to clarify its efficacy. Second, the technique faces significant practical challenges to its users. Virtual reality devices are sometimes presented as a way to solve these practical challenges, but currently fall short of delivering on that promise. We address both issues in this paper. First, we argue that an embodied, enactive approach to memory can better help us understand the effectiveness of the memory palace. Second, we present design recommendations for a virtual memory palace. Our theoretical proposal and design recommendations contribute to solving both problems and provide reasons for preferring an embodied, enactive account over an information-processing treatment of the memory palace.

This chapter is published, in a slightly modified form, as: Peeters, A. & Segundo-Ortin, M. (2019). Misplacing memories? An enactive approach to the virtual memory palace. *Consciousness and Cognition*, *76*, 102834. Anco Peeters is the main author of this chapter, having authored the first three sections, introduction, and conclusion, and taking the lead on structuring the argument. Anco Peeters and Miguel Segundo Ortin co-wrote the fourth section. Both authors contributed to polishing the text. Miguel Segundo Ortin permits the inclusion of the chapter in the present thesis.

Miguel Segundo Ortin

October 15, 2019

Chapter 2

Misplacing memories in virtual reality

"The best aid to clearness of memory consists in orderly arrangement"

Cicero, *De oratore*, 2.86.354, trans. E.W.B. Sutton

Though the memory palace technique, a mnemonic making clever use of places and images, is enjoying newfound attention by researchers on virtual reality (VR), its use goes back centuries. According to one famous story, Giordano Bruno, a Napolitan philosopher and influential memory palace master, earned himself an accusation of plagiarism while presenting at Oxford in 1583. Apparently, one attentive Oxford don did not appreciate that Bruno, in a top-off-the-head lecture, recited long text passages from a contemporary scholar without a reference (Rowland, 2008, p. 146). Bruno's mnemonic use was careless, yet his memory feats remain impressive. Cognitive scientists have been trying to make use of the memory palace more accessible through visualising the technique's places and images in VR, but their efforts have so far yielded underwhelming results. In this paper, we address the issues surrounding recent attempts at operationalizing the memory palace through VR and we present a new and improved way of understanding the technique. Our proposal is inspired both by going back to the technique's roots and by insights from embodied, enactive cognitive science and should help towards solving the issues mentioned.

The main problem with mastering the memory palace technique is the time and effort involved. The technique takes long-term practice, in a suitable environment, and requires creative imagination. This explains

why, given the strength of the technique, its use in education and training practices is not more prevalent. To increase accessibility of the memory palace, researchers have attempted to operationalize its use through VR devices. So far, it has proved hard to gain similar levels of remembering with the use of such devices when compared to traditional mnemonics.

To make steps towards solving this issue, we propose to consider the difficulties in the translation of the memory palace into VR against the background of the so-called 'pragmatic turn' in cognitive science. The pragmatic turn signals a move towards conceiving of cognition as dynamic, embodied and enactive and away from cognition as information-processing (Engel, 2010; Engel et al., 2013). Reframing how we think about the cognitive underpinnings of memory will help in the design of the virtual memory palace.

What is the advantage of examining the memory palace from the perspective of embodied, enacted cognition? We provide two related incentives. The first stems from the observation that current cognitivist investigations into the workings of the technique, which are based on the information-processing paradigm, have not shed sufficient light on why it is so powerful, as we will elaborate in the next section.[1] This opens the door to the consideration of an alternative paradigm. The second and related reason is that the memory palace, because it leans heavily on memory scaffolding through environmental resources, calls for a cognitive framework which places the role of the body in the environment front and centre.

Keeping in mind the pragmatic turn, our paper develops as follows. In Section 2.1, we will examine current cognitivist approaches to the memory palace technique and show how they are unable to explain its dynamics, concluding that there is, as we call it, an Explanation Problem. Following this, we will argue in Section 2.2 that current attempts to operationalize the memory palace in virtual reality fall short, because they depend on cognitivist understandings of the technique. Call this the Operationalization

[1] We take inspiration from a recent critique on symbolic interfacing with augmented reality devices. Raja and Calvo (2017) argue that instead of programming augmented reality glasses (like Google Glass) to navigate spaces using symbols and icons like arrows and text (cf. Clark, 2003, p. 52), such devices would instead function better if they leveraged their user's sensorimotor capacities through changes in brightness. Froese (2014) provides a similar, generalized critique of symbolic interfaces.

Problem. Because addressing the Operationalization Problem first requires addressing the Explanation Problem, we turn to the latter in Section 2.3, where we argue that an enactive account of the memory palace captures the technique better than its cognitivist rivals. This sets the stage for Section 2.4, in which we address the Operationalization Problem by presenting design recommendations for designers of virtual memory palaces based on our proposed enactive account. In doing so, we will rely on influential theories in embodied cognition, such as ecological psychology (Gibson, 1979; Chemero, 2009). We conclude with some considerations on the application of the virtual memory palace in educational settings and for future lines of research.

2.1 The memory palace in cognitive science

Much of our understanding of the memory palace is derived from historical sources. In her titular and seminal book on the art of memory, historian Frances Yates (1966) develops a now classic account of the memory palace. Drawing on instructions by Roman rhetoricians like Cicero and their further development by Bruno, she explains that the memory palace strategy rests on two pillars: *loci* (places) and images.[2]

A *locus* is characterised as part of a spacious environment with distinct features. Classic examples of such environments include large and varied buildings with decorations inside, such as churches and cathedrals. Environmental parts which qualify as loci are usually those that stand out when one would take a familiar route through the environment, such as a gargoyle statue at the entrance, or a niche under a window. Loci and images play a role during both the learning and the recalling phase of the technique. In the learning phase, one moves through the building (preferably physically) and has to imagine placing images of that which has to be remembered at specific locations in and around the building. Then, during the recalling phase, one imagines moving through the building and gets triggered by the images positioned there to reconstruct the memory. It is advised to use vivid and personally resonating images for maximum recall-effect.

[2] In fact, the technique is often called *method of loci* (MOL), though this is a bit of a misnomer as it puts undue focus on the first of the two pillars.

To illustrate the use of the technique and draw out some important aspects, let us imagine the following. While applying the technique to a talk on robot ethics, I choose the Sydney Opera House as my locus of choice. During the learning phase, I physically move around the Sydney Opera House. Initially I imagine a porter at the entrance who holds a copy of Isaac Asimov's *I, Robot*. Moving on, I approach the Opera House's wardrobe, where I imagine Aristotle arguing with Immanuel Kant while Jeremy Bentham hands his head to a robot attending the cloakroom. I continue to move around and create and place images for every part of my talk. When I am ready to present the talk I enter the recalling phase. During that phase, Asimov's book serves to remind me that I need to start my talk by presenting the three laws of robotics, both as an introductory 'hook' and to mark them as a starting point in robot ethics. The image of Aristotle and Kant arguing triggers me to say that virtue ethics and deontology might have something to say on robotics, though both theories are not dominant in current discussions. This is where the image of Bentham comes in, as it cues me to say that utilitarianism is currently the dominant theory in debates on robot ethics. The vividness and personal quality of the images will help me remember, and placing them at specific positions in the locus will help me to order my recollection. The use of personal imagery in combination with the scaffolding of memories through environmental cues are the defining features of the memory palace.

An impressive study by Eleanor Maguire and colleagues (2002), on the functional and neurological differences between normal and high-performing memorizers, shows that the memory palace technique is much alive today. Of the high-performing memorizers, drawn from a pool of participants in the World Memory Championships, 90% report using the technique for some or even all of their tasks. The goal of the study was to capture the possible causes that could differentiate superior memorizers from normal ones. As expected, the superior memorizers performed significantly better in tests on both working and long-term verbal memory. No differences in terms of general intellect or brain structure between the two groups were found. However, functional brain-imaging showed that the superior memorizers, in contrast with the controls, had consistent higher activation levels in the medial parietal cortex, retrosplenial cortex, and the right posterior hippocampus. These regions are "known

to be important for memory, and are implicated in spatial memory and navigation" (p. 93). Unsurprisingly, these brain areas showed increased activity during the learning phase of the task. Thus, Maguire and colleagues conclude that mnemonics like the memory palace, which they defined as "strategies for encoding information with the sole purpose of making it more memorable" (p. 93), constitute the main explanatory cause for the performative difference between superior and normal memorizers. The memory palace technique provides the "top participants of the annual World Memory Championships ... the ability to memorize hundreds of words, digits, or other abstract information units" and is therefore called the "most prominent mnemonic technique" (Dresler et al., 2017, p. 1227).

As of yet, there is no single explanation for why the memory palace technique is so effective. There is nonetheless a suspicion that the "additional motor imagery aspect is likely the reason the method of loci has been found to be particularly effective—a connection that has not been previously made" (Madan & Singhal, 2012, p. 220). This in contrast to other memory strategies which often solely depend on visual imagery. However, it is unclear exactly why motor imagery in combination with visual imagery would explain the effectiveness of the memory palace as a cognitive technique.

Moving further down these lines of thought, Martin Dresler and colleagues (2017) hypothesize that with the memory palace technique "abstract and unrelated information units are transformed into concrete and related information patterns that can more easily be processed by memory-related brain structures, such as the hippocampus" (p. 1232). But what does it mean to say that "concrete and related information patterns" are more easily processed by brain structures? What does the memory palace technique do which transforms a random deck of playing cards from "abstract and unrelated information units" into "concrete and related information patterns"? This transformation seems to presuppose two types of information: abstract and concrete. Are there such different kinds, and, if so, why is concrete information more easily digested? We will take a closer look at this issue in Section 2.3.

The relation between, on the one hand, Yates' account of the memory palace as deeply dependent on both the environment for structure and the individual for creating images, and, on the other, the information-

processing paradigm of the previously discussed experiments, remains underdeveloped. The support of the environment is, in this paradigm, defined as the ordering of information units which are processed by a cognizer's brain. But that this reordering allows for more efficient information processing is, at best, in need of further explanation, or, as we will argue, a fundamentally flawed approach to the understanding of the technique. Let us call this issue the Explanation Problem.

2.2 The virtual memory palace

The Explanation Problem, as we argue in the next section, lies at the root of why efforts at making the memory palace accessible through VR devices, are not yielding results comparable to traditional memory palace practice. Why is there a need for a 'virtual memory palace'? Memory theorists have observed that the "primary flaw of mnemonics is that effective use often requires extensive practice" (Madan, 2014, p. 3). And, specifically in the case of the memory palace, not only does it take practice but it also takes time to familiarize oneself with a large and spacious building and to translate what one wants to remember into images which can then be placed in and around that building. Moreover, the learning phase can be extra problematic for someone who may not always have ready access to a locus that fits the described purpose. Large, easily accessible buildings fit for practice are after all not always available when one wants to, for instance, practise and memorize a talk. Furthermore, the creating-and-placing-the-images phase of the memory palace technique depends on having a creative imagination to come up with evocative pictures which translate to whatever it is one would like to remember. So while the memory palace is acknowledged as a powerful mnemonic technique, potential users are often hesitant to go through the effort of learning it.

Virtual reality technologies might hold an answer to the previously outlined challenges. Virtual environments can be tailor-made for and readily accessible to the memorizer and, when a database of (personalizable) three-dimensional models is provided, the creation of a fitting image for a certain idea in a speech would not be so complicated. The time it takes to practise the mnemotic would also decrease when a virtual environment is available, as there is no need to physically travel to a suitable environment

or spend time conjuring up an imagined one. In the words of Thomas Jund, Antonio Capobianco and Frédéric Larue (2016), given "its intrinsic spatial nature, VR seems to offer the perfect technology devices to implement ... [the memory palace]. Not only [does] it allow ... immersive exploration of any given architectural environment, but it also provides rich sensory cues (spatial contiguity, optic flow, self-directed navigation)" (p. 533). In theory, virtual reality seems to be made, as it were, for the memory palace technique.

Early research on investigating the memory palace through the lens of virtual reality aimed to establish whether virtual environments can support the memory palace technique as well as conventional, physical environments do. In an initial and exploratory study, Eric Fassbender and Wolfgang Heiden (2006) found that participants who interacted with a virtual environment through the use of a personal computer and desktop monitor remembered images from that virtual environment better than words from a sheet of paper. This study is limited because different types of items were compared – images with words – in a within-subject design without randomisation, and there was no between-subject comparison that compared the virtual memory palace to a conventional one. Furthermore, more immersive interfaces than a desktop computer monitor are now available for a consumer market. Higher levels of immersion in virtual environments, specifically in terms of field of vision, improve performance on memorization (Ragan, Sowndararajan, Kopper & Bowman, 2010). This shows that it is preferable to use, for example, a head-mounted display (HMD), rather than a desktop computer monitor to interface with a virtual environment (see also Huttner & Robra-Bissantz, 2016).

In a foundational study on the virtual memory palace, Eric Legge and colleagues (2012) addressed the question of whether the memory palace technique works as well with aid from a virtual environment as from a physical one. In order to test this, the experimenters assigned participants to three groups: a traditional memory palace group, a virtual memory palace group, and a control group. All participants first practised on a memory task, recalling lists of words, then moved through a virtual environment, and finally performed another memory task similar to the first. The first two groups were asked to use the memory palace on the second task, with the former imagining familiar place like their home and

the latter imagining the virtual environment just before encountered. The third group were not given a specific strategy to use.

The results of Legge and colleagues' (2012) research confirm that a virtual environment does not perform worse than a conventional space. However, at least two critical remarks can be made about the study. First, the participants in the study were not present in the virtual environment during the learning phase of the memory task. Instead, they were shown the virtual environment for five minutes and those in the virtual memory palace group were then asked to use their memory of the virtual space for their task. Hence, the study does not speak of how effective the memory palace technique could be when the whole learning phase is performed in a virtual environment. Second, the level of immersion in the virtual environment was again quite low: the environment was shown on a desktop monitor and movement occurred by means of mouse and keyboard. This runs counter to the theory of the conventional memory palace where an active, bodily involvement from the memorizer, in terms of navigation and image placement in the loci, is supposed.

In an effort to make the virtual memory palace a more immediate and immersive experience, Jund et al. (2016) present a study in which participants engaged with a virtual environment by means of an HMD that provided a stereoscopic image. Three types of environments were presented. In the first, participants were sequentially and briefly shown items for remembering in the same frontal virtual position, without spatial cues. In the second, participants were sequentially shown items to remember, with each item briefly appearing next to the location of the previous item in the virtual environment. No further spatial cues were given. The first two conditions were categorised as 'egocentric'. In the third, participants were guided through a virtual apartment with nine different rooms. In this third condition, categorised as 'allocentric', participants used a passive navigation technique: they were moved along a preprogrammed path and could only move forward by pressing a key. Jund and colleagues were surprised to find that the egocentric conditions resulted in better memorization than the allocentric condition. In a follow-up experiment, they adjusted the third condition and found that participants performed significantly better when using a virtual environment of a familiar building. We do not think this result is surprising as per Yates' (1966) suggestion that the memorizer

should use a building which is intimately familiar to them. In the next sections, we argue that an essential cognitive part of the memory palace technique is the training of a cognizer's memory in such a way that it allows for effortless re-imagining of the building in question. In a manner of speaking, such a memorizer would carry the building with them, though we emphasise this should not be understood representationally. However, even with this performance improvement on the allocentric condition, Jund and colleagues found that this condition still did worse than the egocentric ones.

We point out two likely aspects which may help explain the poorer results in the allocentric condition when compared to the egocentric ones in the study by Jund et al. (2016). Both figure in the learning phase of the memorization process. First, the participants could only indicate the moment of movement, upon which they were passively moved along a pre-set path. Second, the participants were presented with images, rather than given the opportunity to create and actively place images in the virtual environment. Both aspects signify the passive relation of the participant to the employed environment and this runs counter to the active anchoring as described by Yates (1966). Jund and colleagues seem to agree, at least on the first point, when they conclude that "the navigation technique and sensory cues associated with displacement might be of primary importance when it comes to use spatial information to support memorization" (p. 537). A new and improved experimental design would be required to determine whether our proposal holds merit, though, and we will provide a design suggestion in Section 2.4.

In a study designed to determine whether immersive HMD interfaces perform better in memory tasks than desktop computer monitors, Eric Krokos, Catherine Plaisant and Amitabh Varshney (2019) take an embodied and embedded approach to the virtual memory palace. Unsurprisingly, they found that the increased immersion of an HMD allows for better memory recall than a traditional desktop monitor. Of even more interest are the peripheral observations they made regarding the manner of interaction between participants and virtual environment. About a third of the participants "mentioned that they actively used the virtual memory palace setup by associating the information relative to their own body" (p. 10). The authors further remark on the previously discussed tension

between active and passive movement through an environment. They refer to Barbara Brooks (1999), who found that active movement allows for more accurate familiarisation with an environment when compared to passive movement. However, as the same study also concluded that the manner of movement, namely whether it was active or passive, had no influence on the recall of items or their positions in the environment, Krokos, Plaisant and Varshney suggest that "memory was only enhanced for those aspects of the environment that were interacted with directly – particularly the environment which was navigated" (p. 4). It should further be noted that Brooks' findings are based on a traditional desktop computer monitor interface with mouse and keyboard, and it would be of interest to redo his experiment with an HMD and direct, haptic interaction of the participants.

Until now, research on the virtual memory palace has presented the memorizer as a somewhat passive participant. We think the observations made by Krokos, Plaisant and Varshney (2019), on the role of the body in (virtual) environments, merit closer attention if we are to properly understand the memory palace technique and develop appropriate interfaces for it – like, for example, via haptic controllers. In line with Krokos and colleagues, we propose to have future experiments assign free movement to the memory palace users in VR. But we suggest departing from this experiment in two ways. First, the images used for testing in the virtual environment were pre-given, while masters of the memory palace emphasise using personalized imagery for stronger memory evocation. Second, the order of images in the virtual environment was signalled by symbols (the numbers 1, 2, and so on). In Section 2.4, we present a way of using lighting to direct the user's attention in virtual environments, to move away from symbolic cues.

With this review of current developments in the field of the virtual memory palace in place, we conclude there is currently no conclusive answer to the question of whether a fully immersive approach, with head-mount display and haptic controllers, can perform as well as (or even better than) conventional memory palace techniques. This means that there is a need for research which compares memory performance of memory palace practitioners both using a conventional memory palace and a virtual

one.[3] It would furthermore be interesting to compare the performance of memory palace practitioners not using a virtual memory palace with ordinary subjects using a virtual memory palace, to establish whether VR operationalization of the memory palace is on par with traditional usage. Based on our interpretation of Yates (1966) in relation to our review of current scientific approaches to the virtual memory palace, we surmise that new research needs to take at least the following into account. First, such an approach needs to investigate what sensory and navigational cues can best support the memory palace. Second, the role of the body in virtual environments needs to be more pronounced than it has been, specifically in terms of how the body is virtually reproduced and whether a haptic interface to the architecture of the locus and the placement of images can enhance the technique. Third, this approach has to promote the active engagement of the memorizer in navigation, choice of loci, and choice of image. Let us call this challenge, to integrate embodied implementations of the memory palace in VR, the Operationalization Problem. It should be clear by now that addressing the Operationalization Problem requires rethinking our cognitive approach to the memory palace, in other words, it requires addressing the Explanation Problem.

2.3 Addressing the Explanation Problem

In addressing the Explanation Problem, we consider two different and competing frameworks which put the embodiment and embeddedness of the cognizer in a larger environment centre stage: extended functionalism and enactivism. In what follows, we connect the memory palace to broader debates on embodied, extended cognition and evaluate the two proposals just mentioned. Our conclusion is that the enactive approach offers more

[3] Another way to look at virtual memory palaces is through the lens of augmented reality devices. In a study performed at the MIT Media Lab, Rosello, Exposito and Maes (2016) present the NeverMind application. NeverMind is designed to run on spectacles or 'smart glasses' which can project images on existing physical locations in the field of vision of the user. The preliminary study found that images projected along a route with NeverMind were better remembered than a list of words on a paper. While definitely an interesting approach, NeverMind still depends on having an appropriate physical environment available. Furthermore, it suffers from the same passive involvement of participants as the studies of Legge et al. (2012) and Jund et al. (2016). As such, it falls beyond the scope of our paper.

powerful resources to account for the effectiveness of the memory palace than its functionalist competitor.

Our examination starts from recent suggestions made in cognitive anthropology and philosophy of mind. Cognitive anthropologist Edwin Hutchins (2005, p. 1564) recounts that the memory palace makes

> opportunistic use of space. The spatial relations of the landmarks do not contribute any semantic content to the problem. But the landmarks themselves do provide memory cues, and the sequential relations among the landmarks, that were created by mapping a particular shape of motion onto them, is inherited by the set of items to be remembered.

This seemingly supports the idea, outlined in Section 2.1, that smart rearrangement of 'concrete and related information patterns' allows such patterns to be more easily processed. However, understanding Hutchins this way would skirt over a crucial difference between his description and the currently salient idea on the memory palace in neuroscience. Instead of focusing on how information patterns might be picked up by the brain, Hutchins, using terms like 'landmark' and 'motion', rightly emphasizes the role of environmental triggers to cue memories and of bodily movement to help in the ordering of them.

The relevance of environmental resources to thinking about the memory palace has also been emphasised by John Sutton (2007). Using the distinction between engrams, or biological memory, and exograms, or external memory carriers, Sutton interprets the physical environments the memory palace technique relies on – like the Sydney Opera House in our example – as "prostheses" or "internalized exograms." Such prostheses, he adds, should be seen as "structuring supplements which construct and maintain the biological processes which they simultaneously and deeply transform" (p. 27).

We will now consider the contribution of such environmental resources from the perspective of extended functionalism. Extended functionalism aligns with current information-processing accounts that we have discussed in the previous sections and can be traced back to Andy Clark and David Chalmers' (1998) classic paper on the extended mind. In this paper, they question the traditional cognitive boundaries of skin

and skull and argue that mind can sometimes be partly constituted by parts of the environment. Clark and Chalmers argue their point by way of their famous thought experiment about Inga and Otto. Inga and Otto are both looking to visit the Museum of Modern Art while in New York. But whereas Inga uses her biological memory to recall the museum's address, Otto, a sufferer of early-onset Alzheimer's, retrieves it through his notebook. The notebook, Clark and Chalmers argue, plays the same role for Otto that biological memory plays for Inga (p. 13). In it, Otto stores the things he would like to remember and his daily routine depends structurally, not incidentally, on his writing – similar to how Inga depends on her biological memory. It is important to note that this constitution claim is stronger than the trivial claim that mind is (merely) causally affected by the environment.

Early extended mind theorists stressed the idea that physical boundaries do not demarcate the mental and argue for this by way of the so-called parity principle. The idea is that "[i]f, as we confront some task, a part of the world functions as a process which, were it done in the head, we would have no hesitation in recognizing as part of the cognitive process, then that part of the world is (so we claim) part of the cognitive process" (Clark & Chalmers, 1998, p. 8). The parity principle encourages us to think that restraining cognitive processes merely to, for example, the brain, would be a case of misplaced neural chauvinism.

The parity principle is the main reason the extended mind is usually seen as part of the larger cognitive programme of functionalism (Clark, 2008a; Wheeler, 2010b, 2015), roughly the idea that mental states are to be defined and characterized by the job they perform. Focusing on functions, instead of material realizers, opens up the way to think that some cognitive processes can be implemented, at least partly, by elements outside the skull. Therefore, theorists working on functionalism are neutral with respect to the whereabouts of cognition, thus providing a natural home for the extended mind thesis.

So how exactly does the memory palace relate to the extended mind hypothesis? Sutton (2010) proposes that, even though mnemonic devices such as the memory palace are not literal external artefacts, the structures they provide function much like Otto's notebook. In this way, Sutton expands the reach of the initial extended mind hypothesis by arguing it can

capture not only natural and biological objects, but also cultural practices. He therefore concludes that "taking EM [Extended Mind] seriously ... means that we treat such architectures, systems, and practices as both cognitive and extended whether or not they happen to be outside the skin" (p. 209).

Let us then give a tentative account of the memory palace according to an extended functionalist framework. As said previously, mental states are, for the functionalist, to be understood in terms of the job they perform.[4] Extended functionalists cast these jobs in terms of information-processing – recall Inga and Otto and that "the information in the notebook functions just like the information constituting an ordinary non-occurrent belief" (Clark & Chalmers, 1998, p. 13). Biological memory is, on this framework, understood as a process which involves the storing and retrieving of informational content, where this content is "sitting somewhere in memory waiting to be accessed" (Clark & Chalmers, 1998, p. 12). When an event is experienced, some piece of information is stored to be later retrieved when required. It has to be noted, however, that the extended functionalist would emphasize that it "doesn't matter whether the data are stored somewhere inside the biological organism or stored in the external world. What matters is how information is poised for retrieval and for immediate use as and when required" (Clark, 2003, p. 69). In light of this framework, we could understand the memory palace technique as a way of structuring the contents and marking them through image-association. During retrieval, the memorizer recollects the relevant contents while she imagines walking through the palace. The images are encountered, the information they encode picked up, and integrated into that which was to be remembered. On this account of extended memory, remembered contents are conceived of as accessible, objective commodities (see Loader, 2013, p. 167).

This type of canonical, "first wave" (Sutton, 2010) extended cognition thinking seems to come some way in explaining the memory palace. It helps us to think of the memory palace as a cognitive structure which supports the memorizer in placing images in a particular order. However, there are two flaws with the current functionalist explanation. First, though it putatively captures the role the environment plays in the process of encoding and retrieving information, it neglects to explain why the role

[4] For a current and general functionalist account of memory, see Fernández (2018).

of bodily movement in both learning and recall phase of the memory palace is of importance. Second, it is unclear how, on this account, the extra information the memory palace would presumably require being processed during the recall phase, actually helps with remembering.

Some extended functionalists, however, have enriched their account to accommodate the role of the body. Clark (2008a), in advocating extended functionalism, proposes two different takes on the role of the body. On the one hand, there is what he dubs the 'Larger Mechanism Story' (LMS), while, on the other, we find the 'Special Contribution Story' (SC). These two stories are explanatorily competitive in that they each assign a different role to the body in the context of embodied cognition.

On LMS, the body is thought to play a special role in the larger information-processing mechanism. To illustrate, Clark (2008a) compares the mental calculation of a sum by a human with how a snake, called Adder, may slither across the keys of an electronic calculator in such a way as to achieve a similar result. He concludes that in both cases the same cognitive operation is performed. The process of the snake's body moving over the keys is functionally equivalent to whatever activity the brain putatively performs to process the relevant information. Because the calculation of the sum is defined in terms of symbol manipulation, extended functionalists can abstract away from the specific material implementations of the calculation and, as such, consider that the body of the snake is no more special than whatever parts of the brain realize these operations. Clark associates LMS with the general (extended) functionalist agenda.

The story is, unsurprisingly, different for SC. On SC, as advocated by Lawrence Shapiro (2004, 2019), the role of the body is not that of one informational piece of the puzzle among many. Instead, as the name implies, those who adhere to SC advocate that at least some of the contributions the body makes are not reducible to mere informational processes. The implication is that some of an organism's cognitive processes are shaped by the specific features of its body in a way that does not lend itself to an explanation in terms of information-processing. Shapiro specifies that there are at least two ways in which the body may influence cognition: "first, it might generate associations that determine certain cognitive proclivities; second, the body might, via activation of motor plans, facilitate or inhibit

various cognitive processes" (p. 12). Thus, on SC, for the understanding of at least some cognitive processes the consideration of the role of the body is required.

To justify the body's role in shaping cognition, Shapiro (2019) draws on empirical sources. Illustrating the first path of the body's influence, he cites research which shows that right-handers prefer to interact with objects on their right side, and left-handers on their left. The idea is that the increased ease with which people interact with objects on their dominant side informs their concept of "good" or "preferred" (Casasanto, 2009, 2014). How would that human preference for one's dominant hand be translated to LMS with a functional description such that a handless organism would exhibit similar cognitive dispositions? Or, as Shapiro (2019) puts it, should "we expect Adder to prefer objects to its right or its left given that it has no hands?" (p. 11).

Empirical evidence supports the notion that at least certain acts of memorizing depend on a special contribution from the body, and we can divide those into the two pathways distinguished by Shapiro. In terms of the first way, that of association, research in psychology has uncovered the relevance of the context-dependence of memory (Smith & Vela, 2001). One foundational study in this regard showed that divers who memorized material while under water better recalled those materials while being under water, while material learned on dry land was better recalled on land (Godden & Baddeley, 1975; Sutton & Williamson, 2014). In terms of the second way, we can draw on the idea that the activation of motor plans are relevant in acts of memorizing, particularly those acts of memory which involve the unfolding of a sequence. I might, for example, try to remember my PIN code by, physically or imaginatively, moving my fingers in its familiar pattern, or recall the order of the alphabet by mouthing parts of it. Scientific research supports this idea, showing that a specific starting point and reenactment through bodily movements is involved in the recollection of interconnected sequences both in musical parts (Ginsborg & Sloboda, 2007; Leman & Maes, 2014; Chaffin, Demos & Logan, 2016) and dance phrases (Kirsh, 2013; Stevens, Malloch, McKechnie & Steven, 2003). On this account, humming a tune or moving one's foot involves the triggering of the next instance in a sequence, domino-style, by the instantiation of its predecessor.

Contextual relevance and the unfolding of familiar patterns are both distinctive aspects of the memory palace technique. Yates (1966, p. 4) stresses that the strength with which a memory is triggered depends on carefully crafted and intense images. Furthermore, the whole environment of a memory palace may contribute to the act of associative recall, as with the divers underwater. Similarly, the sequence with which the images are encountered at the different loci and, as mentioned previously, the neuroscientific evidence of brain areas normally associated with navigation activating during the technique together point towards the idea that motor plans unfold offline during the recall phase (see Section 2.1). Such relations between the role of the body and the memory palace do not conceive of "the body as playing an information-processing role in cognition" (Shapiro, 2019, p. 9) and so the LMS, as cast in its familiar functionalist garb, is unable to adequately capture the memory palace.

For these reasons, we propose to look at an enactivist theory of mind and memory that is, we argue, better able to explain the special contribution of the body in acts of remembering. Enactivism understands cognition not in terms of the processing of information, but in terms of the participation of an organism in sensorimotor loops of active engagement within the context of a larger environment (Varela et al., 1991; Thompson, 2007). Evan Thompson (2007), one of enactivism's main architects, suggests that remembering is better understood, not as the retrieval of a mental image, but as the reproduction of a person's past experience and that it "could involve emulating earlier sensory experiences and thus reenacting them in a modified way" (p. 291).

Enactivists of a radical stripe have further developed this line of thought, casting remembering as a dynamical, re-creative act. Radical enactivists argue that basic forms of cognition do not involve mental representations (Hutto & Myin, 2013, 2017). In line with this research programme, Daniel Hutto and Anco Peeters (2018) put forward the idea that procedural memory "can be understood as the capacity to reenact embodied procedures – often prompted and supported by patterns of response that are triggered by external phenomena" (p. 105). Rather than depending on the metaphor of memory as the storage and encoding of information, a radically enactivist take on procedural memory "would focus not on access to the contents of a store but on remembering as a

type of action" (Loader, 2013, p. 168). Familiar patterns of response are initiated by internal or external triggers. For example, the remembering of how to prepare a specific meal is triggered by the ingredients and tools which are available to the cook. These familiar patterns involve the activation of trained neural configurations, which, according to context and circumstance, enable specific acts (see Anderson, 2014, 2015). Following a recipe in order to prepare a meal is, on this account, not the retrieval of the stored information on that recipe, but the re-enactment of the different steps required to make dinner according to external signposts (the onion is glazed) which direct the individual to follow a specific familiar path (lower the fire).

Procedural memory is in current debates commonly characterized as not relying on information-processing (Michaelian, Debus & Perrin, 2018), but enactivism is not limited to accounts of procedural memory per se. Recently, a number of scholars have proposed that episodic memory centrally involves the construction and consideration of possible past episodes through simulative imagining (Gerrans & Kennett, 2010; De Brigard, 2014; Michaelian, 2016). Such proposals assume that episodic acts of remembering, because of their simulative nature, necessarily involve representational content. Memory theorist Kourken Michaelian (2016), who agrees that understanding procedural memory need not depend on positing representational content, claims, by contrast, that appealing to contents in the case of episodic memory is essential. The reason is that episodic memory is declarative: it is available to consciousness and affects behaviour (pp. 27–28). However, why not allow that episodic memories, like the remembering of a conversation last week, is an act of, perhaps imperfect, simulative reconstruction through which a proposition with the content of that conversation is formed and available to consciousness? That this is indicative of current thinking about memory is shown by Michaelian, who recently argued that radically enactive remembering aligns well with an emerging tendency in discussions of philosophical of memory which cast remembering as non-contentful (Michaelian & Sant'Anna, 2019). In following Hutto and Peeters (2018a), we see no need to assume that all acts of remembering through simulative re-enactment depend on the manipulation

of informational content. We maintain that acts of memory, such as using the memory palace, can be explained in a non-representational way.[5]

Applying an enactive account of memory to the memory palace then leads us to the following theory. In the remembering phase, the memorizer would either walk or imagine walking through an appropriate environment, such as the Sydney Opera House, with which she has become intimately familiar through active, bodily exploration. The order of the loci in the environment ensures that they are sequentially triggered during the recall phase, but it is up to the memorizer to ensure that the loci are then associated with the images to be remembered. During the recall phase, the memorizer will use her imagination to sequentially reconstruct the environment through the familiar triggers. For example, in the case of the Sydney Opera House, she would not remember the Opera House as a whole. Instead, she would reconstruct the relevant features while she images walking through it, letting the triggers guide her. Because of the learned association with the images, these images will spring to mind and can then be used by the memorizer to reconstruct whatever it is she would like to remember. The previously discussed findings by Dresler et al. (2017), on the structural rearrangement of neural networks for users of the memory palace, can then be reinterpreted as the construction of a network which enables the triggering sequence – in essence, a well-practised user of the memory palace carries the triggers of its loci with her. The user of the memory palace is, on the enactive account, not picking up information

[5] One might rightly ask how reconstructive or simulative processes of enactive remembering unfold if they are not based on information storage. While this is an important issue that deserves further elaboration, it is also an open question that needs to be addressed by enactive approaches to memory in general. A proper discussion of this unfortunately falls outside the scope of the current paper. As a tentative proposal, we suggest that enactive remembering involving the previously mentioned processes depend on the sensorimotor activation of familiar patterns. To illustrate, we refer to how artificial neural networks can be trained to generate images (Goodfellow et al., 2014). Such networks do not store specific pixels, but depend on adjusting the signalling strength between nodes during training. After training they may then activate areas on a pre-given (digital) canvas and thus generate an image. Similarly, a person, with an adult, developed brain, may be triggered to think about the Sydney Opera House because of a word read or a sound heard. This trigger may generate, through many intermediary steps, partial images of white, rounded domes against the background of water. It may even be that this person will use her consciousness to help herself generating the memory, for instance, by asking herself "Are the distinctive white shells of the Sydney Opera House spread across two or three separate parts of the building?" Naturally, this is a gross simplification, but it serves as an initial step towards developing a robust enactive account of remembering.

but reconstructing something resembling that which she was supposed to originally remember.

Though enactive remembering seems well suited to explain the role of embodiment in the memory palace while those bodily engagements are not straightforwardly intelligible in information-processing terms, extended functionalists may counter with an adjustment to their theory. In a striking experiment, Wendy Mackay and colleagues (1998) investigated the adaptation of new electronic air-strips at an airtraffic station. In the late 1990s, traditional paper-made strips contained information about speed and direction of incoming airplanes and were used as an integral tool in the safe control of air traffic around Paris. Researchers were tasked to investigate how the use of such strips could be improved or even replaced with electronic devices. Initial trials with replacing the paper-based system with a computer-based one met with resistance by the traffic controllers. Advocating extended functionalism, Michael Wheeler (2010b) observes that, from the perspective of an engineer, "one is inclined to focus, naturally enough, on the information carried by these strips. But this is not the only contribution of the strips." (p. 33). It turns out that the strips were used in ways beyond merely carrying information. For example, they may be held in the hand as a reminder, placed at an angle to indicate two planes on a potential collision course, or, supported by the use of a strip-holding board, afford the signaling of important flight movements through body language. Wheeler's analysis is worth quoting in full:

> From a practical perspective, this recognition of the non-informational contribution of the flight strips is far from idle. The testimonial evidence suggests that a number of previous attempts to introduce new computer technology into air-traffic control may ultimately have been rejected as unworkable by the controllers precisely because *the proposed replacement systems attempted to reproduce the straightforwardly informational aspects of the flight strips while ignoring the extra factors.* (Wheeler, 2010b, p. 33, emphasis added.)

Wheeler concludes that "nothing about this story undermines the extended functionalist line" (p. 33). This implies that the extended functionalist's story either needs elaboration on the differences between 'straight-

foward informational aspects' (like the writing on the strips) and material informational factors (like the orientation of the strips), or that it need not be an information-processing story exclusively. Extended functionalism, as advanced by Wheeler, can thus allow for the materiality of artefacts, such as flight strips, to implement cognitive states as well, because it is neutral with respect to what cognitive states are made of.

Allowing extended functionalism to go beyond merely information-processing by recognizing the material roles artefacts play, looks like a promising move to give a functionalist account of the memory palace. As Wheeler admits, though, his proposal needs further analysis. We see two paths which the extended functionalist could take. The first one is to develop an account which explains the interplay between the informational processing of memories and the role the body plays when walking, imaginatively or not, through the memory palace. Recall that cognitive scientists currently explain the memory palace technique as somehow transforming abstract information units into concrete information patterns. The functionalist needs to provide an explanation of these types of information, explaining whether or not these are different kinds of information, and how transformations between the two take shape. While perhaps not logically impossible, this path seems to lead to conceptually murky waters (Hutto & Myin, 2013, Ch. 4).

A second path for the extended functionalist is to get rid of informational talk altogether and lean on an embodied approach to the memory palace which is entirely non-representationalist. This might seem like a radical move to some philosophers, but it looks like Wheeler is opening the door to that possibility. And a brief look at the history of functionalism provides ground for supporting this move. As Gualtiero Piccinini (2010) argues, functionalism in its purest form is merely the metaphysical claim that cognitive processes are to be understood as structural organizations with input and output relations (see also Putnam, 1967).[6] It seems that an extended functionalist account of the memory palace based on bodily engagement and not on information processing, is a possibility.

[6] Not all functionalists might agree with the claim that pure functionalism is merely a metaphysical claim. However, my aim here is not to present some kind of essential feature of functionalism, but to trace the genealogy of the extended functionalist line back to its most general shape, like Piccinini (2010) does.

Yet, if functionalists surrender their commitment to the information-processing framework, then what difference is left between extended functionalist and enactivist approaches when it comes to explaining the memory palace? It seems the functionalist's metaphysical account would, to the extent to which they could explain techniques such as the memory palace in terms of bodily engagement, collapse into their competitor theories on enactivism (see Hutto, Peeters & Segundo-Ortin, 2017). Elucidating the implications of this collapse lies beyond our current argument, but we would be interested to hear what an adapted extended functionalist story would offer that our enactivist story does not.

The extended functionalist, then, has two options. Either develop an information processing account that is not only able to explain how the body plays its role in the memory palace, but also the transformation of abstract into concrete information (whatever that may be). Or, she could surrender her commitment to information-processing altogether and adopt a fully embodied and non-representational account of the memory palace which basically collapses into an enactive account. In any case, the functionalist is currently not in a position to explain the memory palace while the enactivist is not trapped in a similar dilemma. We conclude that thinking about the memory palace from an enactivist perspective is therefore the better option.

We submit that a radically enactive account of memory, which depends on cues and triggers for re-enactment, may act as a clarifying lens through which to look at mnemonic techniques that centrally involve interaction between a person and their environment, such as the memory palace, whether virtual, imagined or otherwise. As we have seen in the previous section, cognitive scientists currently explain the memory palace in terms of information encoding and retrieval, which leads to virtual memory palaces in which the memorizer is a passive participant with only a superficially strong connection to the used locus. Such operationalizations are better served by an enactivist approach which explains why a multimodal memorization technique that heavily involves visualisation, active involvement of a body with an environment, and the reconstruction of memories is more efficient than learning words from a list. The latter mnemonic after all, provides less triggers and cues with which to rebuild memorized items, while the former builds upon such resources and abilit-

ies for reconstruction which are already in place. Our next step, then, is to determine which resources and abilities a virtual memory palace needs to work on.

2.4 Addressing the Operationalization Problem

How can VR technologies support the practice of the memory palace technique? We propose that VR can support the practice of the memory palace in at least two ways. First, by supporting the user with a virtual memory palace inspired by recent discussions in cognitive science, thus both relieving the user of the need to go to a familiar, physical building to practise and making sure that the virtual environment evokes those sensorimotor interactions which resemble traditional memory palace usage. Second, by enhancing the memory palace technique by actually going beyond that which is feasible through traditional methods, for example by sharing virtual memory palaces with other users or by supplying the user with visual cues to improve memorisation. These two notions form the inspiration for the following operationalization proposal.

As said earlier, deciding on how best to support the memory palace technique in VR depends on one's answer to the Explanation Problem. In contrast to existing operationalizations of the memory palace we argue for an enactive and re-creative account of remembering. If this argument strikes true, it has implications for the operationalization of the memory palace in VR. Specifically, it means that such operationalizations need to be rethought through the perspective of an embodied cognizer which takes the movement within and active engagement with her (virtual) environment seriously and moves away from the idea that using the memory palace is merely a way of reordering and picking up information. In what follows, we propose that adopting an enactive take on memory will support the practice of the virtual memory palace and that it may help to solve the Operationalization Problem of current designs. We do so by giving concrete design recommendations based on this enactive approach.

To move away from the information processing model of the virtual memory palace, the role of the memorizer needs to be recast from passive observer to active participant. In order to do so, we will single out two aspects of current memory palace operationalizations and translate them

into active, body-engaging modes of interaction: movement of the user, and the creation and placement of images. As discussed in the previous section, this translation has to keep in mind the unfolding of sequences through the activation of motor plans in acts of memory. This requires active participation of the body.

Regarding the first aspect, instead of the user being moved through a virtual space passively, we propose that any VR operationalization of the memory palace ought to depart from the idea that the user is actively moving herself through an environment – say a virtual apartment or cathedral. This is not only in line with Yates' (1966) account, which posits the individual moving through the space and engaging with sensori-navigational cues as an essential part of the technique, but also with the two main insights gleaned from current memory research as discussed in Section 2.1. The first is that, at least in some situations, the activation of motor plans supports remembering – recall the examples on PIN codes, music, and dance from the previous section. We have argued that the memory palace is of a similar kind to those examples and thus involves motor activation. Second, neuroscientific evidence supports the idea that brain areas associated with spatial navigation are involved in the use of the memory palace. As such, we think approaches where users are either passively moved or there is no movement at all, do not support the optimal unfolding of a memory sequence.

Moving on to the second point, the placement of images, we present a similar line of reasoning. Active participation of the body in the placement of images in a virtual space would mean that the user should be able to do two things. First, she should be able to either choose or, preferably, create personalised images which may represent parts of that which she wants to memorize. A database in the virtual space, where images can be stored and retrieved, can support the user friendliness and re-use and easy adjustment of images. Second, the user should then be able to place those images in distinct locations in the virtual memory palace. Virtual reality devices with hand-held controllers that can mimic regular hand movements seem especially suited for these use-cases.

Now that we have discussed how the memory palace technique could be translated to VR, by using insights from an enactive approach to cognition to improve movement and image placement, we will present ways

Figure 2.1: Imaginary virtual memory palace with a suggested *locus* highlighted.

of potentially enhancing the virtual memory palace. Is it possible to go beyond the technique's traditional limitations? And if so, how?

One way in which to take advantage of computer technology is to highlight features of the virtual environment in such a way as to support the user's needs. In this, we take inspiration from work done by Vicente Raja and Paco Calvo (2017), who propose a way of looking at augmented reality based on ecological psychology (Gibson, 1979). In discussing navigational apps, such as Google Maps, they argue that instead of overloading a user by presenting yet more symbolic information on a screen, for example, by showing a top down map with arrows and numbers, certain pathways might be emphasised more subtly. For instance, one can imagine a user wearing smart glasses which brighten those areas that the user should go, and darken areas the user should avoid. This nudges a user into the destination she wants to go to. Similarly, we suggest, parts of one's virtual memory palace can be highlighted during the learning phase if they offer a memorable location to carry an image associated with part of what one wants to remember (see Figure 2.1). Or, also during learning, when unfolding the sequence of the memory the next part of the sequence in the virtual space that a user needs to go to can be brightened, visually, as the

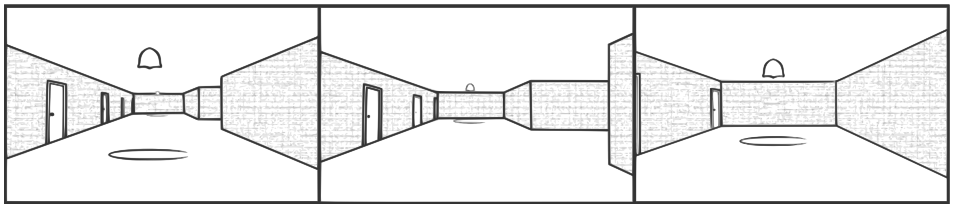

Figure 2.2: The walls in these consecutive images expand in a process of optic expansion.

next space to move to. So instead of overloading the user with symbolic information, a virtual environment might support memory performance by highlighting the relevant affordances this environment offers to the user (Stoffregen, Bardy & Mantel, 2006).

A second way of enhancing the virtual memory palace concerns what we dub 'sensorimotor realism.' Note that realism here should not be understood in its common, digitalized meaning: as the photo-realistic replication of images and textures. Contrary to this, perhaps intuitive, idea, there is empirical evidence which suggests that familiar sensorimotor interaction in virtual environments contributes more to the immersion of the memorizer in her memory palace than high-resolution imagery (Fink, Foo & Warren, 2009). Sensorimotor interaction in VR further seems to improve one's sense of agency, in the sense of experiencing control over one's actions and their consequences (Kong, He & Wei, 2017), which ties in nicely with and supports the previously discussed active bodily participation.

By *sensorimotor* realism we mean that a VR device involving movement needs to replicate the kind of sensory patterns we experience when we move in real life. To illustrate, think of what occurs when you approach a wall. As you approach the wall, you see how the texture gradients of the wall radiate from the centre of your visual field, causing the wall to expand from the perspective of the perceiver (see Figure 2.2). This is commonly described by saying that optic flow is centrifugal in the direction of locomotion (Chemero, 2009, p. 124). The rate at which optic flow expands is lawfully correlated to the speed to which we move towards the object – the wall in this case. By saying that a virtual environment must be sensorimotor realistic we mean that it must echo the sensorimotor experience we are used to in real life. The optic flow generated while moving towards

an object in the virtual environment ought to be the same as the one we get when we do so in real life. Otherwise, our experience of moving through the virtual space will feel odd and unpleasant (Bubka, Bonato & Palmisano, 2008), and it will require us to take extra effort to get attuned to the sensorimotor contingencies of the virtual environment. Ensuring sensorimotor realism will thus add to the immersiveness of the virtual memory palace.

Incorporating active bodily participation lies at the heart of our proposal for operationalizing the virtual memory palace. For the translation of the memory palace to VR, we argued that this requires the user to take control in the virtual environment. For potentially enhancing the virtual memory palace, we proposed to make use of sensorimotor guidance that makes optimal use of the type of interactions the user is already familiar with.

2.5 New horizons for memory research

Considering the memory palace from an embodied, enactive perspective, in line with the pragmatic turn in cognitive science, helps in understanding why current operationalizations of the technique in VR leave much to be desired. Such operationalizations focus on supporting the picking up of information by the user, but we have argued that this does not capture what is at the core of the technique.

Instead, we presented design recommendations for improving the virtual memory palace, focusing on embodied cognition and affordances. Smart use of VR devices could make the learning of the memory palace more accessible and increase the usage of one of the most powerful methods of remembering on offer. Our design recommendations are ready for implementation. If their adaptation yields better results than current operationalizations, this will have both practical and philosophical implications. To start with the latter: if virtual memory palaces based on our enactive proposal work well *outside* of the head, it would provide a good reason, by way of abduction, to re-evaluate what is going *inside* the head. By way of a reversed parity principle, the enactivist research programme would have provided an impressive case in point in terms of understanding the

underpinnings of memory, placing the ball squarely in the functionalist park.

The practical implications, if our proposal holds true, lie in making the power of the memory palace more accessible and their advantages are obvious. Special attention should be given to its potential use in educational settings (Putnam, 2015). We predict that using VR devices to support learning through the memory palace can greatly enhance learning experiences (in line with: Mäkelä & Löytönen, 2017; Heersmink, 2018). Not only that, but activities which are traditionally seen as boring, like the rote learning of words from a foreign language, would potentially become a lot more fun because of the engaged, bodily interaction. Furthermore, in classroom settings, both teachers and students can benefit from the shared experience which VR will allow. Unlike in the traditional technique, teachers would be able to participate in and give feedback on how their students utilize the memory palace.

Our proposal, though grounded on available empirical data, requires more experimentation. Not only to test whether the hypothesized design recommendations will improve the use of the memory palace, but also to investigate aspects of the techniques that were hereto hard or impossible to investigate. The sharing of the same loci, as described in the previous paragraph is one aspect, but this could be generalized to the investigation of loci which are not necessarily environmental landmarks as traditionally imagined. For example, how will moving objects like animals or other persons affect the technique? What about videos? Virtual realities allow for plenty of creative freedom and the memory palace is a worthy candidate for testing the limits of that freedom with respect to successful memory strategies.

Abstract

We propose that virtue ethics can be used to address ethical issues central to discussions about sex robots. In particular, we argue virtue ethics is well equipped to focus on the implications of sex robots for human moral character. Our evaluation develops in four steps. First, we present virtue ethics as a suitable framework for the evaluation of human–robot relationships. Second, we show the advantages of our virtue ethical account of sex robots by comparing it to current instrumentalist approaches, showing how the former better captures the reciprocal interaction between robots and their users. Third, we examine how a virtue ethical analysis of intimate human–robot relationships could inspire the design of robots that support the cultivation of virtues. We suggest that a sex robot which is equipped with a consent-module could support the cultivation of compassion when used in supervised, therapeutic scenarios. Fourth, we discuss the ethical implications of our analysis for user autonomy and responsibility.

This chapter is published, in a slightly modified form, as: Peeters, A. & Haselager, P. (2019). Designing virtuous sex robots. *International Journal of Social Robotics*, 1–12. Online first publication. Anco Peeters is the main author of this chapter, having authored the first three sections, introduction, and conclusion, and taking the lead on structuring the argument. Pim Haselager wrote a draft of the fourth section and Anco Peeters then contributed to it. Both authors contributed to polishing the text. Pim Haselager permits the inclusion of the chapter in the present thesis. Footnote 3 was added to link to scientific developments presented after the publication of the current chapter.

Pim Haselager

November 2, 2019

Chapter 3

Designing virtuous sex robots

> *"We need an ethics that does not stare obsessively at the issue of whether a given technology is morally acceptable but that looks at the quality of life that is lived with technology."*
>
> Peter-Paul Verbeek (2011, p. 156)

Some may find it hard to come to grips with sex robots. Yet recent events, like the 2015 *Campaign Against Sex Robots* in the UK, the 2017 publication of John Danaher and Neil McArthur's volume on the ethical and societal implications of robot sex, and the fourth incarnation of the *International Conference on Love and Sex with Robots*, show that this topic has captured the public's eye and provokes serious academic debate. A recent report by the *Foundation for Responsible Robotics* (Sharkey, van Wynsberghe, Robbins & Hancock, 2017) calls for a broad and informed societal discussion on intimate robotics, because manufacturers are taking initial steps towards building sex robots. We take up this call by applying virtue ethics to analyse intimate human–robot relationships.

Why should we look at such relationships through the lens of virtue ethics? Virtue ethics is one of the three main ethical theories on offer and distinguishes itself by putting human moral character centre stage – as opposed to the intentions or consequences of actions. Virtue ethics has been discussed in relation to artificial intelligence more generally (Wallach & Allen, 2009; Tonkens, 2012). However, virtue ethics has received relatively little attention in discussions regarding sex with robots, even though sex robots could have a significant impact on their user's moral character. Two main exceptions are Litska Strikwerda (2017), who assesses

arguments against the use of child sex robots, and Robert Sparrow (2017), who suggests that rape representation by robots could encourage the cultivation of vices. Our aims are different, as we will not focus on either child sex robots or robots that play into rape fantasies. Instead, we propose how virtue ethics can be used to contribute to the potential positive aspects of intimate human–robot interactions through the cultivation of virtues, and provide suggestions for the design process of such robots.

We develop our thesis in four steps. First, we present virtue ethics in relation to other ethical theories and argue that, because of its focus on the situatedness of human moral character, virtue ethics is in a better position to assess aspects of intimate human-robot interaction (see also Vallor, 2016, p. 209). Second, we show how our virtue ethical account fares better than current instrumentalist approaches to sex robots, such as those inspired by the seminal and pioneering work of David D. Levy (2007a, 2007b). Such instrumentalist approaches focus too much on the usability aspects of the interaction and, unjustly, frame sex robots as neutral tools. Understanding the interaction with a sex robot as mere consumption insufficiently acknowledges the risk of their influence on how humans think about and act on love and sex. Third, we propose a way to reduce the risks identified by considering how the cultivation of compassion as a virtue may help in practising consent-scenarios in therapeutic settings. This way, we aim to show how, under certain conditions, love and sex with robots might actually help to enhance human behaviour. Fourth, we examine the implications our virtue ethical analysis on intimate human–robot relations may have on our understanding of autonomy and responsibility.

3.1 Virtue ethics and social robotics

Current ethical debates on human–robot interaction are generally not framed in terms of virtues, but in terms of action outcomes or rules to be followed. It strikes us as regrettable that up until now, virtue ethics has received relatively little attention in the literature on social robotics in general, and on intimate human-robot relations in particular (but see Abney, 2012; Gips, 1995). A virtue-ethical analysis can help evaluate how, on the one hand, human agents could make use of love and sex robots in ways that may be judged to be (un)problematic. On the other hand, virtue

ethics may help to clarify how human behaviour and societal views are influenced by the use of such robots and thereby help us to learn more about what it is to be a virtuous person in an intimate relationship. To establish the potential of virtue ethics for the evaluation of intimate human–robot relationships, we will examine aspects of virtue ethics relevant to the current discussion and consider what it has to add compared to other ethical approaches.

Virtue ethics departs from the idea that the cultivation of human character is fundamental to questions of morality. In the Western philosophical tradition, Aristotle's theory of virtue ethics is the most influential and he defines virtue as an excellent trait of character.[1] Such traits, like honesty, courage and compassion, are stable dispositions to reliably act in the right way according to the situation one is in. Aristotle describes a virtue as, in general, the right mean between two extremes (vices). He states that courage, for example, can be described as the mean between recklessness and cowardice (*Nicomachean Ethics*, II.1104a7). Finding the right middle between extremes is a challenging task and approaching that middle often requires extensive practice. In addition to practice, acquiring a virtue is helped by instruction from an exemplary teacher. A virtuous person will have cultivated her character to be disposed to naturally act in the right way in the relevant situation. It should be noted that although virtues are not about singular acts, acting honestly, courageously or compassionately may help a person to become honest, courageous or compassionate. This potential interactive loop, of internalising behaviour by practice and feedback, motivates our interest in applying virtue ethics to intimate human-robot interaction.

Consequentialism and deontology are the two main rival theories to virtue ethics, and they dominate current discussions on the ethics of social robotics. Consequentialism is the ethical doctrine that takes the outcome of an action as fundamental to normative questions. Deontology or duty-based ethics takes the principles motivating an action as central to matters of morality. Operationalization of these frameworks can take different forms. For example, in the case of consequentialism, artificial

[1] Other influential virtue ethical traditions originated with, for example, Confucius or Buddhism. For reasons of space, we shall restrict ourselves to a (neo-)Aristotelian account of virtue, but we suspect that the investigation of other virtue traditions could yield an interesting intercultural approach to the ethics of social robotics. See also Vallor (2016).

agents could be programmed to evaluate the potential costs and benefits of an action (Deng, 2015; Winfield, Blum & Liu, 2014; Sharkey, 2008; Floridi & Sanders, 2004). Or, in the case of deontology, designers may strive to implement top-level moral rules in agents (Danielson, 1992).[2] As consequentialism and deontology provide frameworks that can be translated relatively straightforward into implementation guidelines, they may be attractive from a roboticist's perspective. While we value the contributions of consequentialist and deontological approaches to the literature on robot ethics, we think that there are ethical issues which virtue ethics is in a better position to address. Such issues include how, in the words of Shannon Vallor (2016), advances in social robots are "shaping human habits, skills, and traits of character for the better, or for worse" (p. 211). Importantly, this insight supports the idea that robots are not neutral instruments, but that they may influence the way we think and act. We side, therefore, with other researchers who recognize that virtue ethics can be a fruitful framework for AI and robotics (Abney, 2012, p. 37).

There are at least three ways in which virtues (and vices) might play a role in social robotics. First, we may consider which virtues are or ought to be involved on the human side of robot design. For instance, is it desirable that a roboticist exhibits unbiasedness and inclusiveness when designing a robot? Second, robots may nudge users towards virtuous (or vicious) behaviour. An exercise robot, for example, can encourage proper exercise and discipline by giving positive feedback to its user. Third, robots may exhibit virtues (and vices) through their own behaviour. This can be illustrated by the Sociable Trash Box robot developed at Michio Okada's lab at Toyohashi University of Technology (Yamaji, Miyake, Yoshiike, De Silva & Okada, 2011): these robots exhibit helpfulness and politeness through their vocalisations and bowing behaviour when they collaborate with humans to dispose of trash (see Figure 3.1). So one could focus on the virtues of the designer, on the way robot behaviour affects the virtues of a human interacting with it, or on the virtues displayed by the robot, for instance, as an example to be followed or learned from. We will focus on the latter two points, but towards the end discuss their implications for design.

[2] Isaac Asimov's famous laws of robotics, often cited as illustration in the ethics of AI literature, are modelled after deontological formulations of how one ought to act. They brilliantly showcase the inherent tension between deontological robotic directives and the potentially disastrous consequences that strict adherence to these might have.

Figure 3.1: The Sociable Trash Box exhibits helpfulness and politeness when it requests trash and then bows after receiving it. Reprinted by permission from Springer Nature: Springer *International Journal of Social Robotics* (Yamaji, Miyake, Yoshiike, De Silva & Okada, 2011), © 2019.

We think it is likely that the degree of anthropomorphism (Sparrow, 2002, 2016; Cappuccio, Peeters & MacDonald, 2019; Björling, Rose, Davidson, Ren & Wong, 2019) will play an important role for especially the second and third topics. This needs to be further investigated, but for the purposes of this chapter we will discuss robots that tend towards the anthropomorphic rather than the more functional end – like conventional sex toys – of the anthropomorphism spectrum.

In relation to the third aspect, some have said that virtues might be difficult, or even intractable, to implement in a robot. This idea is motivated by the complexity of giving general, context-independent definitions of specific virtues and because an implementation of a virtue like honesty "requires an algorithm for determining whether any given action is honestly performed" (Allen, Varner & Zinser, 2000, p. 258). Although we acknowledge the specific implementation challenges that virtue ethics brings, we think these challenges can be addressed by looking at the underlying mistaken assumption that virtues need to be implemented top-down into the robot. Analogous to how humans learn to be virtuous not by being told what to do but by example, implementing virtues into the design of social robots can take a similar situational approach. For this reason, it

has been argued that the "virtue-based approach to ethics, especially that of Aristotle, seems to resonate well with the [...] connectionist approach to AI. Both seem to emphasize the immediate, the perceptual, the non-symbolic. Both emphasize development by training rather than by the teaching of abstract theory" (Gips, 1995, p. 249). This resemblance, we suggest, can help inspire the implementation of virtues in modern-day robots. The use of machine learning with artificial neural networks may be a way of avoiding the need to write an algorithm that specifies what action needs to be taken when. Virtues that depend on, for example, recognizing emotions in a human and require an emotional response can be implemented by training a neural network on selected input – say, by analysing videos of previously screened empathic responses made by humans (as done by Janssen et al., 2013; Güçlütürk et al., 2017). Through machine learning, robots could similarly learn to mimic certain behaviours that we might consider displays of virtue, such as a light touch on the shoulder to express sympathy.[3] The challenging research question here would be how to operationalize this kind of training so that the robot learns from human teachers. Such implementations are not trivial, but they need not be intractable either.

Two potential points of critique need to be addressed before moving on. The first critique has been voiced by robot ethicist Robert Sparrow (2017), who argues that sex robots could encourage vicious behaviour, while at the same time maintaining that he finds it hard to imagine sex robots could promote virtue. He proposes that if people own sex robots, they can live out whatever fantasies they have on the robots – even rape. He argues that repeated fantasizing and repeated exercise of potential representations of rape will influence one's character to become more vicious. Though we agree with Sparrow's premise that this development is problematic and deserves careful consideration, we disagree with the conclusion drawn. While rape representation might be facilitated by sex robots, this does not mean that the production of such robots need always be ethically inimical. Let us assume that rape-play between two consenting

[3] After the publication of the present chapter, research by Senft, Lemaignan, Baxter, Bartlett and Belpaeme (2019) has shown how it is possible to teach robots human-like social behaviour through mimicry and machine learning, with the authors specifically mentioning application of their research in therapeutic scenarios.

adults is not necessarily morally wrong.[4] What is potentially morally wrong in acting out this scenario, is that it might normalize the associated repeated behaviour outside of a consensual context – the cultivation of a vice. This could lead to unwanted degrading behaviour or generalization to other contexts involving human-human interaction. The same risk of inappropriate generalization applies to the scenario of the human–robot interaction. In the case of humans, this means that careful and continuous communication about what is allowed and what is not is crucial: the partners have to trust and respect each other in order to safely play out the fantasy and stay aware of the fact that it is a fantasy. Might a similar approach be possible to intimate human-robot interactions? We submit that there are ways to involve consent in the case of intimate human-robot interaction aimed to prevent the risk Sparrow is drawing attention to, without condemning the manufacture and use of sex robots in principle.[5] It would require us to rethink sex education and the role sex robots can play in this, which we do in Section 3.3. Interestingly, if one accepts that sex robots may cultivate vices in humans, it seems possible that such robots potentially also cultivate virtues.[6]

A second issue that needs addressing is a more general critique against virtue ethics. It has been argued that virtue ethics as an ethical theory is "elitist and overly demanding and, consequently, it is claimed that the virtuous life plausibly could prove unattainable" (Fröding, 2011, p. 223). Why propose such a demanding ethical theory for framing human-robot interaction? First, because virtue ethics can do justice to an assumption we make, namely that intimate, sexual relations between humans and robots

[4] It is worth noting that on Sparrow's account one will have to bite the bullet and say that rape-play by consenting adults is morally wrong as well. Not everyone will be willing to accept this implication.

[5] Obviously, the consent provided by a robot does not amount to legally binding consent, just like the rape of a robot would not constitute legal rape, for the simple reason that a robot is not a legal person and not a sentient being. Hence, we are discussing here the implications of a robot behaving in a certain way, not necessarily implying the existence of human-like cognitive, emotional states or identical legal status.

[6] Sparrow (2017) finds it "much less plausible that sustaining kind and loving relationships with robots can be sufficient to make us virtuous" (p. 473). He acknowledges, however, that such a claim needs to be supported by an argument as to why virtues are to be held against a standard different from vices and that this is a topic for further discussion. We do not share his intuition, though we agree with his latter point and would furthermore like to add that more empirical data on how human–robot interaction influences human behaviour is needed – which is one of the motivations for the proposal in Section 3.3 of the present chapter.

	Consequentialism	*Deontology*	*Virtue ethics*	*Instrumentalism*
Fundamental concept	Action outcomes	Moral rule	Virtue	Instrumental use
Concept applied	Obtaining consent maximizes well-being for both parties.	Obtaining consent is in accordance with the rule: "Do unto others as you would be done by."	Obtaining consent is compassionate and respectful.	Obtaining consent is not necessary, unless required for obtaining satisfaction.

Table 3.2: Suppose we compare the multiple approaches in a hypothetical scenario where sexual consent is negotiated, verbal or otherwise, between two human partners. This table aims to show how such a scenario can be analysed in the different ways discussed in the present chapter. This rough distinction should not be taken to mean that, for example, consequentialism cannot talk about virtues. What distinguishes the different approaches is which concept they take to be central.

should be understood as bi-directional. In this context, bi-directional means that humans design robots, while the general availability of such robots in turn may influence human practice of and ideas on intimacy and love. In contrast, current ways of thinking about intimate human-robot relations often depart from an instrumental and unidirectional assumption. Such rival accounts understand these relations as the usage of tools by humans and see any influence that robots may have on humans as value-neutral. They are focused on the human perspective and therefore lose sight of important potential ethical implications of human-robot interaction, as we will argue in Section 3.2 and as illustrated in Table 3.2. Our assumption is in line with current developments in cognitive science and philosophy of technology, which suggest that the cognitive and moral dimensions of artefact interaction need to be understood from a distributed perspective that puts equal emphasis on agent and environment (Varela et al., 1991; Verbeek, 2011; Coeckelbergh, 2012; Di Paolo et al., 2017).

Another and possibly even more exciting reason to engage with virtue ethics, is that thinking about virtues in relation to robots might actually help to make virtuous behaviour more attainable. This might be done through the habit-reinforcing guidance of humans by robots designed to promote virtuous behaviour: either by robots nudging human behaviour directly or by robots exhibiting virtues themselves.

3.2 Contra instrumentalist accounts

Recent discussions on intimate human–robot relations are often informed by the work of David D. Levy (2007a, 2007b). Levy argues that humans will have physically realistic, human-like sex with robots and feel deep emotions for and even fall in love with them. Although we laud the pioneering work Levy has done to open up sex and love with robots for serious academic discussion, we argue that his framework fails to properly account for the ethical and social implications involved.

Regarding sex, Levy suggests that, physically speaking, realistic human-like sex with robots will be possible in the near future. Though Levy paints a colourful history of the development of sex technologies, discussion of this is not of prime importance for our argument and we will not examine it further. For the present discussion, we will assume that the physical aspects of these robots can be worked out more or less along the lines which Levy describes. Interestingly, Levy goes so far as to say that "robot sex could become better for many people than sex with humans, as robots surpass human sexual technique and become capable of satisfying everyone's sexual needs" (D. Levy, 2007a, p. 249).

Regarding emotions and love, Levy suggests that it is possible that humans can be attracted to and even fall in love with robots. Without going into too much unnecessary detail, his argument proceeds in four steps. First, Levy lists what causes attraction of humans to each other. Second, he considers how affective relationships between humans and pets develop, and, third, how such relationships develop between humans and their *virtual* pets. Fourth and finally, Levy applies his findings to human–robot relationships.

Through a careful examination of feelings of bonding and attraction in humans, Levy comes to the conclusion that humans will likely develop similar feelings of bonding and attraction for robots. A large role in this narrative is reserved for the human tendency to anthropomorphize artefacts (see Breazeal, 2002; Sparrow, 2002). He submits that "each and every one of the main factors that psychologists have found to be the major causes for humans falling in love with humans, can almost equally apply to humans falling in love with robots" (D. Levy, 2007a, p. 128). It seems that there are no major hindrances for humans to, at some point in the

future, fall in love with their robot. We can, in principle, agree, with this conclusion and it furthermore looks like recent preliminary empirical evidence supports it (Scheutz & Arnold, 2017).

Obstacles on the path towards the use of love and sex robots are deemed by Levy to be of a merely practical nature. The robots described are presented as taking care and recognizing the needs of their human partner – in terms of the feelings of bonding and attraction he listed earlier. On several occasions (D. Levy, 2012, 2007a, pp. 219, 233) Levy compares sex with a robot to masturbation, and uses that comparison as a reason why robot-sex would prevent cheating on one's partner (p. 234) – like in the case of soldiers on a long-term mission. Moreover, Levy describes this perspective on sex as a kind of "consumption" (D. Levy, 2007a, p. 242). It is for this reason that we characterize accounts such as Levy's as 'instrumentalist.' Love and sex robots, on such accounts, are merely tools to be used or products to be consumed. However, we suggest that such an instrumentalist perspective could lead to practices that provide cause for concern. Also, we are not convinced that a purely instrumentalist use of sex robots would make many people "better balanced human beings" (p. 240).

A first concern is that framing robot-sex as consumption underestimates the potential impact the acceptation of love and sex robots will have on the way love and sex are perceived. Consider a world where your "robot will arrive from the factory with these parameters set as you specified, but it will always be possible to ask for more ardour, more passion, or less, according to your mood and energy level. At some point it will not even be necessary to ask, because your robot will, through its relationship with you, have learned to read your moods and desires and to act accordingly" (D. Levy, 2007a, p. 129).

Why would people, when such partners are available, be content with any kind of relationship, emotional or sexual, that would not adhere to this standard of perfection? Access to these robots would make it tempting to view relationships as essentially one-directional need-catering and effortless, especially perhaps for adolescents who grow up with such access. This is not how love and sex at present needs to be or even generally is conceived, and it goes deeply against the conception of a relationship as existing between two or more equal persons. Seeing humanoid robots

capable of emotional and sexual interaction as tools is like being in a relationship with a slave. There lies an important question at the core of this issue, specifically on whether there are ways of considering the relationship between human and robot that are not slave-like. However, this falls outside the scope of the current chapter (though for a beginning of an answer to this question, see Cappuccio et al., 2019). In any case, this comparison illustrates the extent to which Levy's framework is unidirectional, which is further exemplified by his comparison of robot-sex with masturbation. Masturbation, at least generally speaking, is a solitary enterprise, and does not reflect the reciprocal interaction that characterizes a typical sex encounter between two partners.[7] Precisely because robot-sex does not amount to either masturbation or sex between consenting adults, one needs to address its particular ethical implications.

The second worry is that the instrumentalist approach allows for downplaying the risk of addiction inherent in interacting with robots that can perceive and immediately cater to their partner's every need. Consider how Levy describes that "robots will be programmable never to fall out of love with their human, and they will be able to ensure that their human never falls out of love with them" and "your robot's emotion detection system will continuously monitor the level of your affection for it, and as that level drops, your robot will experiment with changes in behaviour until its appeal to you has reverted to normal" (D. Levy, 2007a, p. 118). This sounds like the perfect gambling machine, which constantly updates its rules according to its user's desires – though these robots are potentially far more addictive than any currently existing gambling machine. We think this issue is insufficiently addressed by instrumentalist approaches such as Levy's, because, if one thinks of robots as merely neutral tools, as he does, then any risk of addiction rests solely on the shoulders of the user and not on a robot or its designers. However, it is an open question whether this is how robot-sex will be experienced by human users (or their significant others). Rather, we suggest that robots are not merely neutral tools.

A convincing argument in this regard is provided by Peter-Paul Verbeek (2011), who argues that for instance an obstetric ultrasound is not merely a neutral tool, a 'looking glass' into the womb. Its use raises im-

[7] This also illustrates that robot-sex is not or need not always be wrong. This would be as extravagant a claim as the suggestion that masturbation is always wrong.

portant ethical questions, like "What will we do when it looks like our unborn child has Down syndrome?" or social pressure such as "Why did you decide to let the child [with Down syndrome] be born, given that you knew and you could have avoided it?", or more general societal questions like "Is it desirable that ultrasonography leads to a rise of abortions because of less severe defects like a harelip?" (Verbeek, 2011, p. 27). This shows that the use of obstetric ultrasound influences our moral domain. It is naive to think that using technologies would not shape our behaviour and societal practices. Instead, it is better to think about this shaping of behaviour while designing technology. Similarly, instead of seeing robots as neutral tools, we should acknowledge that, for instance, robots may evoke more emotions in us than other tools do, as Matthias Scheutz (2012) suggests. More importantly perhaps, the design and use of intimate robots presuppose or establish certain practices concerning 'appropriate intimacy.' At the very least, these practices and their underlying assumptions should be elucidated.

Two conclusions can be drawn from the above account. First, humans and technologies should not be seen as separately existing entities, with technology providing neutral products for human consumption. Secondly, ethical analyses are not based on pre-given ideas or criteria, but need to re-evaluate how human-artefact interaction may be influenced or radically changed by new technologies. This means that stakeholders participating in the design of technologies have a responsibility both in considering how their products will shape human behaviour and reflecting on the ethical issues that may arise with the use of their product.

On this view, designers are "practical ethicists, using matter rather than ideas as a medium of morality" (Verbeek, 2011, p. 90). In this framework there is room for the moral aspects of technologies in a pragmatic context, without it becoming a 'thou shalt not'-like ethics. A virtue-ethical approach is exactly what the topic of intimate relations with robots needs, because interacting with a robot as an artificial partner is, even more so than with a regular artefact, a relationship which intimately shapes our own dispositional behaviour and societal views as well. On first sight, Levy seems open to a more interactive view when he refers to Sherry Turkle, taking up her line of thought in saying that he "is certain that robots will transform human notions" including "notions of love and sexuality" (D.

Levy, 2007a, p. 15). The way Levy discusses situatedness resonates with the notions that humans and technologies should not be seen as strictly separate entities and that certain concepts are not pre-given but arise out of interaction between humans and artefacts. Does that mean Levy has successfully anticipated critique along the lines we have set out? It does not.

Although Levy seems sensitive to the two notions mentioned, in practice it is merely a lip-service to interactive human–technology approaches. His instrumentalist treatment of human–robot relations deals with humans and robots in terms of isolated atoms with only a one-way connection between them, from user to robot, without any consideration of the larger reciprocal interactive effects on behaviour and social practices. He does not analyse robot-sex in terms of the structures and situatedness he earlier described. Any instrumentalist framework will focus on the human, subject side of things and portray robots as neutral artefacts to be used. What Levy describes is a trend of an increasing acceptation of robot sex, not how it would actually constitute or change (our conceptions of) sex or intimate relationships. Even if one agrees that masturbation is not cheating – an open question, likely to be influenced by many contextual factors – that does not necessarily mean that having sex with a robot will not be considered as cheating. An intelligent android functions on a distinctively different level of companionship than, say, a vibrator. More dramatically, if instrumentalist thinkers on the one hand argue that an intimate relationship with a robot is possible and imply that these kinds of relationships can be as intense and realistic as intimate relationships between humans, then they should agree that being intimate with such a robot, while in a relationship with someone else, could be construed as cheating. At the very least, one has to concede that robot-sex in such a scenario cannot simply be equated to masturbation. In other words, even assuming that one would find it hard to imagine someone being jealous about one's partner using a vibrator, one could still imagine jealousy plays a role when one's partner engages in sexual activities with a very human-looking and acting robot.[8]

[8] The Swedish science-fiction television drama *Äkta människor* (*Real humans*, 2012) depicts an example of this when the relationship between Therese (Camilla Larsson) and her husband turns sour because he grows jealous of her 'hubot' – a humanoid robot capable of exactly the functions Levy discusses. This depiction is fictional of course, but the force

The analysis we have given shows that instrumentalist approaches may leave crucial ethical considerations unaddressed. Notions of love and sex will be changed by the development of humanlike robots. But how will these notions change? If we can have sex robots which are "always willing, always ready to please and to satisfy, and totally committed" (D. Levy, 2007a, p. 229), what will that do to the way we view relationships? An understanding of robot-sex not as instrumental, neutral use of tools, but as involving a reciprocal interaction between human agents, robots and their designers is required to develop adequate answers to questions such as these. This is where virtue ethics can provide a guide for evaluation of such interactions.

3.3 Consent practice through robots in therapy

In order to investigate how sex robots could make a positive contribution to human moral character, we draw on virtue ethics for ideas on how to cultivate virtues and connect those to insights from current empirical data provided by literature on robotics and psychology. Our aim is to avoid the problem of cultivating vices through repeated unnegotiated practice – such as illustrated by Sparrow. Indeed, well designed robots may create the possibility to actually improve attitudes and behavioural habits regarding sex. First, consider the human–sex robot rape play scenario again. Previously, we argued that what is problematic about this scenario is not the act between consenting adults itself, but the potential normalization of behaviour it could lead to. For instance, the human participant may become accustomed to immediate satisfaction of desires through the use of a human-looking object and might extend the involved behavioural patterns to objectify other humans.

One way of preventing unwanted behavioural patterns is by providing sex robots with a module that can initiate a consent scenario. Like consenting humans, a robot and its human partner will have to communicate carefully about the kind of interaction that will take place and the human will be confronted by the subject-like appearance and the behaviour of the robot. And like in a relationship between humans, this communica-

of the story at least casts doubt on any outright dismissal of the possibility that humans will become jealous of robots.

tion could potentially result in the robot sometimes not consenting and terminating the interaction. Such interaction with a robot might prevent the practice of unidirectional behavioural habits and a resulting increased objectification of other humans.[9] This consideration suggests that the potential psychological and behavioural benefits of a consent-module will make it at least worthy of investigation. One should notice too, however, that a consent-module may negatively affect the potential economic gains of sex-robot producers, a consequence that is not our main concern here. Second, there are potential benefits with respect to sex practice and cultural perception in general in the consent-module, namely in cultivating the virtue of compassion. Though we focus on compassion for the sake of limiting the scope of this case study, other virtues, such as respect, likely ought to play a role in consent-practice as well. We take compassion here as the ability to care for and open up to another person without losing sight of one's own needs and feelings. Virtuous displays of compassion strike the right balance between care for others and for oneself. Compassion can motivate a desire to help others and we take it to be related to, though distinct from, empathy (see Goetz, Keltner & Simon-Thomas, 2010).

A robot equipped with a consent-module could potentially be used to investigate ways of improving consent practice in general. Often, partners communicate their willingness to engage in sex through nonverbal cues (Byers & Heinlein, 1989). Yet, because nonverbal cues can be ambiguous, miscommunication can and does occur (Abbey, 1991b). In response, some governmental institutions have advocated the need for active, verbal consent. The practice of active consent has been met by at least two problems. First, even verbal consent does not necessarily mean that a partner is freely engaging in sex, because, for example, social pressure or substance abuse may be involved (Lim & Roloff, 1999). Second, explicit consent has met with cultural resistance, as men and women generally believe discussing

[9] On the other hand, one might argue, as Sparrow does, that a non-consenting robot could potentially facilitate (the representation of) rape scenarios even more if the human partner ignores the robot's consent. We do not have a solution for that problem here (although, for example, a simple 'complete close-and-shutdown' routine might be an option), but it is a main reason why we later in this chapter suggest to test this kind of human–robot interaction in a therapeutic setting first, as testing under supervision may give us new insights on how to potentially deal with issues such as these. In any case, we are not convinced that this argument is sufficient to not further investigate the potential benefits of consenting robots.

consent decreases the chance that sex will occur (Humphreys, 2004). Still, active consent is seen as a crucial way of combating sexual assault and rape, for example, at college campuses (Abbey, 1991a; Banyard et al., 2007; Borges, Banyard & Moynihan, 2008). There is a need to change perceptions and practice, especially by men (Adams-Curtis & Forbes, 2004), concerning healthy consent and sexual practices. Virtuous sex robots – supervised – might help facilitate a much needed cultural change in this regard by further investigating ways of navigating consent.

The advantage of using sex robots over traditional top-down education is that the robots can provide a kind of embodied training that helps adolescents in negotiating sexual consent. Interaction with a compassion-cultivating sex robot could raise awareness of how these scenarios could play out and alter behaviour through training. A sex robot which not only can practise consent scenarios with a human partner, but which can actually cultivate a virtue like compassion could potentially be used in sex education and therapy. A robot cannot suffer and so any moral harm during education or training will be minimized. It seems to us that compassion is a suitable virtue to be practised using sex robots in sex education and therapy. If successful in clinical trials, such robots can be used to support a change in perception and behaviour of consensual sex on a larger scale, and not just with adolescents.

One might be sceptical as to whether robots can facilitate a dependable long-term change in compassion – both in negative or positive ways. It seems reasonable not to judge this prematurely, as assessing the long-term effects of sexual human-robot interactions requires empirical investigation by sexologists and psychologists. A number of interesting experiments on the influence of social robots on human behaviour in more general terms, have been done in the lab of Nicholas Christakis. In one (virtual) experiment (Shirado & Christakis, 2017), humans were placed into groups which had to perform a task. Unknown to the participants, these groups also contained robot agents. The robotic agents were programmed to make occasional mistakes which adversely influenced group performance. This behaviour led to the human participants who collaborated directly with a robot, to become more flexible in finding solutions that benefited group performance. Similarly, a related experiment (Traeger, Sebo, Jung, Scassellati & Christakis, 2019) reported that humans who collaborated on

a task with robots which made occasional mistakes and acknowledged their mistakes with an apology, became more social, laughing together more often, and more conversational.

The design of virtuous sex robots requires thinking about a setting in which to test and apply them. A case study will give the constraints necessary for the design to be specific and feasible. We further think that building a robot which can operate in long-term intimate relations in general first requires at least building a robot which can operate on a smaller timescale with a specific target audience. Furthermore, it would be necessary to have the support of supervisors – next to the AI researchers which should of course also be involved – that have professional training in psychology or psychiatry. We therefore propose to start with testing virtuous sex robots in a therapeutic setting.

As the specific target audience or participants, we suggest to consider persons who have been diagnosed with a narcissistic personality disorder (NPD) as the common medical understanding of NPD (American Psychiatric Association, 2013a) aligns well with the previously given definition of compassion. We propose to consider NPD patients who are already within a therapeutic setting, as this means that testing can be done in a controlled environment, under supervision of professionals in psychiatry, psychology, and sexology. The robot's design, testing and development beforehand should involve these same professionals, especially regarding the potential effects of a robot's refusal of certain kinds of interaction. The anticipated link with compassion can be found in the latest edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). In it, narcissism is described as a "pervasive pattern of grandiosity, need for admiration, and lack of empathy" (American Psychiatric Association, 2013a). Nine indicators are listed for narcissistic behaviour, of which the third, fifth, and sixth are of special interest for us here. Respectively, those indicators are about the narcissist feeling special, being exploitative in social relations, and lacking empathy. If compassion as a virtue is the golden mean between two extremes, then it seems that the narcissist, who feels better than others and is self-obsessed, is at one extreme of the spectrum.[10]

[10] In the spirit of virtue ethics, one could consider Dependent Personality Disorder (DPD) to be the other extreme on the compassion spectrum (American Psychiatric Association, 2013a):

We would describe this extreme (or vice) as having the tendency to being overly involved with oneself. Hence, training the virtue of empathy and compassion would be most relevant for this focus group. Designing and evaluating a robot aimed at influencing the behaviour of persons is the most prominent, and challenging, task to be set. Though there is a lack of information on successful NPD treatments (Dhawan, Kunik, Oldham & Coverdale, 2010), there is some preliminary evidence that empathic treatments of those with NPD have positive effects (Bender, 2012).

Obviously, operationalizing our proposal requires careful testing before the possibility of actual use in training is even considered, as the care for patients and the safety of those potentially harmed by their conduct is paramount. One potential worry might be, for example, that people with narcissistic tendencies become more proficient in their manipulations. Therefore, professionals involved would need to closely monitor the patients and signal such possible undesired effects. These cautionary words notwithstanding, the potential support of compassionate robots for NPD treatments is in line with the aforementioned preliminary evidence (Bender, 2012) and worth further investigation.

The next step in making the robot ready to teach compassion is by training it to give basic responses to certain kinds of behaviour. As proposed before, this could be done by training it on recordings of how compassionate people respond to different kinds of (inappropriate) behaviour. This means the robot has to recognize at least one extreme on the compassion spectrum in terms of behaviour of its partner, and has to perform behaviour appropriate to what it observes. Figuring out what good identifiers of those extremes are and what responses work best will need to draw heavily on the expertise of the psychiatrists involved.

Compassion is considered here as the virtue which lies between the extremes of only caring about oneself, the narcissist, or of only caring about another person. That means that a robot designed to treat these kinds of disorders should be able to direct behaviour towards the middle

> They are willing to submit to what others want, even if the demands are unreasonable. Their need to maintain an important bond will often result in imbalanced or distorted relationships. They may make extraordinary self-sacrifices or tolerate verbal, physical, or sexual abuse.

It would be interesting to investigate how love and sex robots could be relevant for training and therapy for members of this group as well.

of the spectrum, where there can be a healthy focus on both caring for oneself and caring for others. We suggest that it may be worthwhile to investigate whether and how such behaviours could be influenced by a compassionate robot. If this turns out to have promising results, work can be done on improving the design and expanding the use of such robots for other settings and for other groups of people.

3.4 Implications of virtuous sex robots

We have striven to demonstrate that virtue ethics provides a useful framework for analysing the implications of sex robots, as well as for making recommendations for the design and application of such robots. We consider robot-sex as involving and supporting a reciprocal interaction between human agents and robots instead of as a form of uni-directional instrumental tool use. Applying virtue ethics led us to suggest a consent-module for sex robots that could support the development or strengthening of compassion in supervised, therapeutic scenarios. As such, sex robots may contribute to the cultivation of virtues in humans. However, virtue ethics does come at a price. In addition to its potential of providing an interesting perspective on the issues surrounding sex robots, it may also raise new problems. As an illustration of the latter, we would like to briefly reflect on two implications of implementing a consent-module. Robots saying 'no' towards the human that uses or owns them can lead to at least two related principled problems and one big practical challenge.

First, robots that refuse to comply with the demands or wishes of human beings may obstruct a person's autonomy, for example, as expressed by someone's immediate or long-term desires (see for a field study in the context of service robots for elderly Bedaf, Draper, Gelderblom, Sorell & de Witte, 2016). Second, there is the threat of a responsibility gap. Finally, there is the practical challenge of how to design such a consent-module. We will offer some minor suggestions to address the latter at the end of this section.

We will illustrate the problem of a user's autonomy by considering a simple example in a different context. Imagine a beer robot, a simple system that keeps a stock of beers cooled and that brings one on demand. Obviously, at some point this might result in intoxication of the person

demanding the beer. To what extent should a ('virtuous') beer robot be enabled to refuse the demands for another beer? Even though the consequences of intoxication may be bad for the persons themselves, as long as no one else or no one else's property is hurt, one might conclude that it is an expression of a person's autonomy to keep the beers coming. It is only or at least primarily in the context of negative effects for other persons or legal agents, that one could morally or legally preclude someone from having their wishes gratified. So, on the one hand, the human should be in control, but at some point or in certain contexts it could be legitimate or morally acceptable to limit the amount of control a human may have.

Regarding the responsibility gap, the problem is that when a human instructs a well-functioning robot to do something, and the robot is programmed to refuse to follow the instructions, all kinds of consequences may follow from that refusal for which the human, in essence, cannot or need not be held responsible. This leads to the question: Who would be responsible or accountable for any damages, psychological or physical, that may ensue? Of course, problems regarding the consequences of saying 'no' are not specific to virtue ethics. Rather, they are a consequence of any view that implies that robots under certain conditions should refuse specific instructions. However, this is worth discussing here because our analysis of virtue ethics leads to proposal of a consent-module, and its consequences should be noted. In our brief discussion, we will try to focus as much as possible on the specific nature of the ensuing problems in the context of sex robots.

In order to address these issues of autonomy and responsibility, we suggest considering the principle of 'meaningful human control'. This principle has been discussed in the contexts of military robots and self-driving cars. The principle states that ultimately humans should remain in control and carry (ultimate) responsibility for robot decisions and actions (Article 36, 2015). However, it is far from clear what this principle amounts to in practice, that is, what the requirements are for the robot so that it is capable of enabling this principle. Filippo Santoni de Sio and Jeroen van den Hoven (2018) indicate that humans merely 'being in the loop' or controlling some parameters may be insufficient for meaningful control if other parameters turn out to be more relevant to the robot's use or if the human lacks enough information to appropriately influence the process.

In addition, possessing an adequate psychological capacity for (assessing) appropriate action is required for meaningful control, as is, thirdly, an adequate (legal) framework for assessing responsibility for consequences. Santoni de Sio and van den Hoven then analyse meaningful control in terms of John Fischer and Mark Ravizza's (1998) theory of guidance control. Guidance control is realized when the decisional mechanism leading up to a particular behaviour is "moderately reason-responsive", meaning that in the case of good reasons to act (or not), the agent can understand these reasons and decide to act (or not), at least in several different relevant contexts. Moreover, the decision-making mechanism should be "the agent's own", in the sense that there are no excusing factors such as being manipulated, drugged, or disordered.

This, admittedly brief, consideration of meaningful guidance control provides a criterion that might be useful for the consent-module. It provides ground to think that when a human does not possess sufficient guidance control, or, by robot compliance with human instructions, may lose such control, a robot could be justified in non-compliance. This leads to two questions that need to be answered before a virtuous sex robot can be enabled with a consent-module, allowing it to refuse commands:

1. Is the person giving the current command in a state of meaningful human control?

2. Will complying with the current command lead to a reduction of meaningful human control, such that (1) is no longer the case?

In relation to the first question, the beer robot could make use of relatively reliable physiological measurements (like breath or blood analyses), or behavioural observations (like slurred speech or coordination difficulties). It will be more difficult to figure out which input patterns might engage the consent-module to generate refusals. Here too, the expertise of psychologists and psychiatrists, in relation to NPD for instance, is required. The main suggestion here is that a DSM-5 classified disorder in itself constitutes a reason for at least considering the possibility that the ability to act reasonably and compassionately might be affected, or that sound judgement and behavioural control might be impaired. Practically speaking, it would be relevant to investigate the extent to which data

acquisition methods related to emotion recognition and sexual harassment might apply. Among potential indicators one could think of, for example, the human's lack of allowing turn-taking in communication, tone of voice and body posture, neglect of robotic non-verbal signals of non-interest, and so on (see, e.g., Miranda, Canabal, Portela García & Lopez-Ongil, 2011; Rituerto-González, Mínguez-Sánchez, Gallardo-Antolín & Peláez-Moreno, 2019). As a second step, investigations regarding the applicability of machine learning techniques are relevant (e.g., Fernandes, Cardoso & Astrup, 2018).

The second question points to a difference between the case of the beer robot and the virtuous sex robot. In case of the beer robot, a prediction about the intoxication can be made on the basis of physiological variables. Given certain physiological aspects, the time course of the intoxication can be inferred with reasonable, and legally satisfactory, certainty. An intoxication level close to life-threatening alcohol-poisoning, just to mention a relatively clear case, could result in justifiable robot non-compliance. However, in the case of the virtuous sex robot such a prediction about the consequences of (non-)compliance is not as straightforward. For this reason too, it bears emphasis that we are suggesting the investigation of the consent-module within clinical contexts. Assuming, for the moment, agreement regarding the appropriateness of a robot's non-compliance in certain situations, there is still a further question about how the non-compliance should be put into effect. We just mention a few possibilities here. One option is that a robot may refuse to comply, provide an explanation in terms of its assessment of the potential negative consequences, and provide information aimed at improved self-understanding and self-control. Ideally, this could result in a retraction of the instruction given. Another option may be that the robot refuses and informs a support group of, say, significant others or therapists. A more extreme option would be that the robot refuses and stops functioning altogether, by way of an emergency close-and-shutdown operation. Finally, it is worth noting that we may need to stretch our concepts of autonomy and responsibility beyond the individual and recast them in terms of open-ended and ecological processes (see Clark, 2007). Unfortunately, picking up this topic lies beyond the scope of the present chapter.

Undoubtedly, many other issues and ways of addressing them surround the notion of a consent-module. We have explicated the present ones to emphasize that virtue ethics does not provide easy solutions. Rather, it opens up a research domain in itself, one that comes with its own set of promises and difficulties that will need to be addressed.

3.5 Next design steps

The field of robotics advances rapidly and robot ethics ought to keep up. In the foreseeable future, there will be robots advanced enough to evoke, even if only for a few minutes, the experience in humans that they are interacting with another human being. Unless a ban is implemented (Richardson, 2016), which we do not want to rule out, it is likely that love and sex relationships with robots will be formed. How can we best understand and evaluate such relationships? We have taken some initial steps towards answering this question by arguing that virtue ethics is better suited than instrumentalist approaches to evaluate the subtleties of intimate human-robot relationships. Next steps should involve careful testing and with this in mind we have outlined how testing a consent-module for robots in a therapeutic setting may yield useful insights. Importantly, implications for user autonomy and responsibility should remain in focus of future research.

Some challenges are anticipated. First, the misuse of sex robots could have a lasting impression on an adolescent learning about intimate relationships, but there is also a positive side to developing realistic looking and acting love robots. Such robots could train people how to behave confidently and respectfully in intimate relationships. In a therapeutic setting, such robots could be used to improve empathy or increase self-love in persons with respectively narcissistic or dependent personality disorders.

Another challenge is society's response to sex robots. It is difficult if not impossible to predict how our conceptions of love and sex will change with the introduction of love robots. One risk here is that a potential societal taboo on love and sex with robots would lead to fringe behaviours and scenes, similar to the domain of drugs and prostitution. It is therefore important that the topic of sex-robots, challenging, exciting, or revolting

as it may appear to different parties, remains open for investigation and discussion.

The implications of developing love and sex robots are potentially huge and we have striven to tentatively chart one path, a virtue theoretical approach, within this domain. Advances in other robotic fields, like care robots or military robots, might have analogous implications. In these areas too, we should avoid the mistake of assuming that robots will not change the way we view healthcare and warfare. On the contrary, we need to consider and assess which of these changes would be desirable or should be avoided. In any case, we would do well to avoid the suggestion that all these developments are necessarily bad. We suggest that there is the possibility, worthy to be investigated, that some changes might be for the good. When we realize that the way we design and use such robots is bound to affect us, we can think about ways of improving ourselves through the technology, by careful consideration and monitoring.

Concluding remarks

In the present dissertation, I have contributed to the development of an enactive account of mind–technology interaction. This has been motivated by recent discussions in philosophy of mind & cognition by a growing number of scholars who advocate a pragmatic turn towards cognition as action-oriented and dynamic. The pragmatic turn moves away from accounts that, inspired by the computer metaphor, conceive of cognition as relying on representations and information-processing. Within this debate, functionalism is often seen as the main champion of the latter kind of theories, while enactivism is heralded as part of the former. A special kind of functionalism, one that departs from the extended mind thesis, is currently the dominant theory when it comes to understanding mind and technology. Current theories on enactivism have not yet yielded a mature theory about the specifics of embodied technology engagements and so an examination of mind–technology interaction within the context of the pragmatic turn is a life issue for enactivists. Therefore, if this dissertation is to bear fruit, it has to be shown that its contributions move the enactivist programme closer – or beyond – its functionalist rival and further support the pragmatic turn in cognition.

The first step towards an enactivist account of mind and technology has been the building of a bridge between enactivism and philosophy of technology in Chapter 1. From the field of philosophy of technology, postphenomenology was presented as a potential partner to enactivism by establishing their common ground. This common ground is most clear in their shared assumption that mind is to be understood as co-constituted by both agential and environmental factors. Enactivism and postphenomenology have been shown to be of mutual theoretical benefit through two steps. First, by discussing how the postphenomenological division of the six kinds of human–technology relations can inform enactivist research

on mind–technology interaction. Second, by exposing how enactivism may provide a cognitive underpinning to postphenomenological research.

Following up on the general issues outlined in the first chapter, Chapter 2 considered a concrete case-study for a critical comparison of functionalism and enactivism. In line with the pragmatic turn, this case-study examined how functionalist and enactivist theories fare with respect to operationalizing the memory palace mnemonic in virtual reality. Supported by a critical review of the current empirical literature on memory and mnemonics, I argued that functionalist theories of memory fail to account for the embodied aspects of the memory palace by their assumption that cognition is information-processing. In its stead, I developed an enactive account of the memory palace and offered design recommendations for its use in virtual reality.

Having discussed the phenomenological and cognitive science aspects of enactive technology engagement, Chapters 3 and 4 turned on the ethical aspects of mind–technology interaction. Chapter 3 cleared ground by arguing that virtue ethics is highly relevant when it comes to the discussion of the impact of sex robots. I positioned a virtue ethical analysis of sex robots against an instrumental use, arguing that the latter fails to capture crucial implications of sex robot use on human moral character. The main contribution of this chapter to the literature is its proposal for the potential positive aspects of sex robot use in therapy. This contribution leans on the idea that sex robots, when outfitted with a consent module and, *crucially*, used in supervised therapeutic settings, might contribute to virtue cultivation. The published paper that this chapter is based on has already garnered much societal discussion and informed specialists working in the field of robotics.

In Chapter 4, I picked up the thread left hanging at the end of the previous chapter, namely what the connection between virtue ethics and enactivism is. By using the situationist challenge against virtue ethics as a foil, I argued for an enactive understanding of human moral character. This chapter offered a rebuttal of situationism by dissolving the opposition between moral character, as traditionally constrained to the body, and environmental factors. Additionally, it has provided a proposal for virtue cultivation through 'self-programming'. In doing both, this chapter has clarified slumbering connections between virtue and enaction.

The principle of multiple realizability, often thought to be a major trump card for functionalism, provided the stage for a brief reflection in Chapter 5. Because of its emphasis on the concrete materiality of cognitive acts, enactivism can be thought to be incompatible with the thesis that cognitive processes can be realized in different physical kinds. However, I have shown that, when we move away from assuming that multiple realization must necessarily depend on information-processing and allow for cognition to be realized over parts of the brain, body, and environment, enactivism can be said to be compatible with multiple realization. This result offers enactivism extra support in comparisons with functionalism.

The main contributions of this work were informed by a number of fields and will, potentially, affect them in turn. By drawing on insights provided by philosophy of technology, cognitive science, and virtue ethics I have articulated how enactivism may categorise and understand different human–technology relations, add to the design of technologies, and elucidate the moral issues surrounding embodied technology interaction. This will enable enactivism to better investigate the crucial experiential, scientific, and ethical issues surrounding mind–technology interaction. Not only does this offer theoretical benefits by informing wider discussions about the pragmatic turn in cognitive science. It also offers new insights to engineering research on the embodied aspects of robotics, virtual reality and the ethical issues surrounding those.

Invariably, a number of questions remain open and I shall briefly discuss the main one here. Whereto can functionalism move to meet the challenges raised in the present dissertation? It became clear, on a number of occasions, that through the so-called 'third wave' in extended mind, there might arise an extended functionalism that does not conceive of cognition as the processing of information. So far, the work on third wave extended cognition has been largely preliminary, but we can single out one potential trajectory. In line with the pragmatic turn, the pressure on functionalists to move away from representations and computational mechanisms has steadily increased. Combine this observation with the fact that a growing number of functionalists accept the extended mind thesis, and the question is raised whether functionalism is on a course of convergence with enactivism. Given the fact that many theorists, from both sides, have, on a number of occasions, stated that functionalism

and enactivism are incompatible, a future discussion of this observed convergence would be highly relevant and interesting.

I aimed for this dissertation to make a positive contribution to the vibrant and interesting discussions on mind and technology. Given the initial uptake of the published papers in this work, I am carefully optimistic that this aim was at least partially realized. It is my hope that the present work empowers us to better understand mind–technology interactions, and, therefore, also ourselves.

Bibliography

Papers by Hilary Putnam which were reprinted in his collected *Philosophical Papers* are referenced by their original publication year, yet any page numbers in the running text refer to the reprinted collection.

Abbey, A. (1991a). Acquaintance rape and alcohol consumption on college campuses: How are they linked? *Journal of American College Health*, *39*(4), 165–169.

———. (1991b). Misperception as an antecedent of acquaintance rape: A consequence of ambiguity in communication between men and women. In A. Parrot & L. Bechhofer (Eds.), *Acquaintance rape: The hidden crime* (pp. 96–111). New York: Academic press.

Abney, K. (2012). Robotics, ethical theory, and metaethics: A guide for the perplexed. In P. Lin, K. Abney & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 35–52). Cambridge, MA: MIT Press.

Achterhuis, H. (Ed.). (2001). *American philosophy of technology: The empirical turn* (R. P. Crease, Trans.). Bloomington: Indiana University Press.

Adams, F. & Aizawa, K. (2001). The bounds of cognition. *Philosophical Psychology*, *14*(1), 43–64.

———. (2008). *The bounds of cognition*. Oxford: Basil Blackwell.

Adams-Curtis, L. E. & Forbes, G. B. (2004). College women's experiences of sexual coercion. *Trauma, Violence, & Abuse*, *5*(2), 91–122.

Alfano, M. (2014). What are the bearers of virtues? In H. Sarkissian & J. C. Wright (Eds.), *Advances in experimental moral psychology* (pp. 73–90). London: Bloomsbury.

Alfano, M. & Skorburg, J. A. (2017). The extended and embedded character hypothesis. In J. Kiverstein (Ed.), *The Routledge handbook of philosophy of the social mind* (pp. 465–478). Oxford: Routledge.

Allen, C., Varner, G. & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, *12*(3), 251–261.

American Psychiatric Association. (2013a). Personality disorders. In *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Association.

———. (2013b). Substance-related and addictive disorders. In *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Association.

Anderson, M. L. (2014). *After phrenology: Neural reuse and the interactive brain*. Cambridge, MA: MIT Press.

———. (2015). Mining the brain for a new taxonomy of the mind. *Philosophy Compass*, *10*(1), 68–77.

Arnau, E., Estany, A., González del Solar, R. & Sturm, T. (2014). The extended cognition thesis: Its significance for the philosophy of (cognitive) science. *Philosophical Psychology*, *27*(1), 1–18.

Article 36. (2015, April 9). Killing by machine: Key issues for understanding meaningful human control. URL: http://www.article36.org/autonomous-weapons/killing-by-machine-key-issues-for-understanding-meaningful-human-control/.

Aydin, C. (2012). Het uiterlijk van het innerlijk: Extended mind, technologie en de binnen-buiten scheiding. *Tijdschrift voor Filosofie*, *74*(4), 701–728.

———. (2015). The artifactual mind: Overcoming the 'inside–outside' dualism in the extended mind thesis and recognizing the technological dimension of cognition. *Phenomenology and the Cognitive Sciences*, *14*(1), 73–94.

Aydin, C., González Woge, M. & Verbeek, P.-P. (2019). Technological environmentality: Conceptualizing technology as a mediating milieu. *Philosophy & Technology*, *32*(2), 321–338.

Bach-y-rita, P. (1983). Tactile vision substitution: Past and future. *International Journal of Neuroscience*, *19*(1–4), 29–36.

Bach-y-rita, P. & Kercel, S. (2002). Sensory substitution and augmentation: Incorporating humans-in-the-loop. *Intellectica*, *2*(35), 287–297.

Baker, L. R. (2009). Persons and the extended mind thesis. *Zygon*, *44*(3), 642–658.

———. (2013). Technology and the future of persons. *The Monist*, *96*(1), 37–53.

Banyard, V. L., Ward, S., Cohn, E. S., Plante, E. G., Moorhead, C. & Walsh, W. (2007). Unwanted sexual contact on campus: A comparison of women's and men's experiences. *Violence and Victims*, *22*(1), 52–70.

Barandiaran, X. E. & Di Paolo, E. A. (2014). A genealogical map of the concept of *habit*. *Frontiers in Human Neuroscience*, *8*(522), 1–7.

Baumeister, R. F. & Masicampo, E. J. (2010). Conscious thought is for facilitating social and cultural interactions: How mental simulations serve the animal-culture interface. *Psychological Review*, *117*(3), 945–971.

Baumeister, R. F., Masicampo, E. J. & Vohs, K. D. (2011). Do conscious thoughts cause behavior? *Annual Review of Psychology*, *62*, 331–361.

Bedaf, S., Draper, H., Gelderblom, G.-J., Sorell, T. & de Witte, L. (2016). Can a service robot which supports independent living of older people disobey a command? The views of older people, informal carers and professional caregivers on the acceptability of robots. *International Journal of Social Robotics*, *8*(3), 409–420.

Bender, D. S. (2012). Mirror, mirror on the wall: Reflecting on narcissism. *Journal of Clinical Psychology*, *68*(8), 877–885.

Björling, E. A., Rose, E., Davidson, A., Ren, R. & Wong, D. (2019). Can we keep him forever? Teens' engagement and desire for emotional connection with a social robot. *International Journal of Social Robotics*. Online first publication.

Borges, A. M., Banyard, V. L. & Moynihan, M. M. (2008). Clarifying consent: Primary prevention of sexual assault on a college campus. *Journal of Prevention & Intervention in the Community*, *36*(1-2), 75–88.

Botvinick, M. & Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature*, *391*(6669), 756.

Brancazio, N. (2019). Gender and the senses of agency. *Phenomenology and the Cognitive Sciences*, *18*(2), 425–440.

Breazeal, C. L. (2002). *Designing sociable robots*. Cambridge, MA: MIT Press.

Brewer, J. A. & Potenza, M. N. (2008). The neurobiology and genetics of impulse control disorders: Relationships to drug addictions. *Biochemical Pharmacology*, *75*, 63–75.

Brook, A. (2009). Introduction: Philosophy in and philosophy of cognitive science. *Topics in Cognitive Science*, *1*(2), 216–230.

Brooks, B. M. (1999). The specificity of memory enhancement during interaction with a virtual environment. *Memory*, *7*(1), 65–78.

Bubka, A., Bonato, F. & Palmisano, S. (2008). Expanding and contracting optic-flow patterns and vection. *Perception*, *37*(5), 704–711.

Byers, E. S. & Heinlein, L. (1989). Predicting initiations and refusals of sexual activities in married and cohabiting couples. *Journal of Sex Research*, *26*, 210–231.

Cappuccio, M. L., Peeters, A. & MacDonald, W. D. (2019). Sympathy for Dolores: Moral consideration for robots based on virtue and recognition. *Philosophy & Technology*, 1–23. Online first publication.

Carr, D. (1991). *Educating the virtues: An essay on the philosophical psychology of moral development and education*. London: Routledge.

Casasanto, D. (2009). Embodiment of abstract concepts: Good and bad in right- and left-handers. *Journal of Experimental Psychology: General*, *138*, 351–367.

———. (2014). Body relativity. In L. Shapiro (Ed.), *The Routledge handbook of embodied cognition* (pp. 108–117). Oxford: Routledge.

Castelvecchi, D. (2019). Black hole pictured for first time — in spectacular detail. *Nature*, *568*(7752), 284–285.

Chaffin, R., Demos, A. P. & Logan, T. (2016). Performing from memory. In S. Hallam, I. Cross & M. Thaut (Eds.), *Oxford handbook of music psychology* (pp. 559–571). Oxford: Oxford University Press.

Chalmers, D. (2017). The virtual and the real. *Disputatio*, *9*(46), 309–352.

Chemero, A. (2009). *Radical embodied cognitive science*. Cambridge, MA: MIT Press.

Chemero, A. & Silberstein, M. (2008). After the philosophy of mind: Replacing scholasticism with science. *Philosophy of Science*, *75*(1), 1–27.

Churchland, P. M. (2005). Functionalism at forty: A critical retrospective. *Journal of Philosophy*, *102*(1), 33–50.

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

———. (2007). Soft selves and ecological control. In D. Ross, D. Spurrett, H. Kincaid & G. L. Stephens (Eds.), *Distributed cognition and the will: Individual volition and social context* (pp. 101–122). Cambridge, MA: MIT Press.

———. (2008a). Pressing the flesh: A tension in the study of the embodied, embedded mind? *Philosophy and Phenomenological Research*, *76*(1), 37–59.

———. (2008b). *Supersizing the mind: Embodiment, action, and cognitive extension*. Oxford: Oxford University Press.

———. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford: Oxford University Press.

Clark, A. & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7–19.

Coeckelbergh, M. (2012). *Growing moral relations: Critique of moral status ascription*. Palgrave.

Coghlan, S., Vetere, F., Waycott, J. & Neves, B. B. (2019). Could social robots make us kinder or crueller to humans and animals? *International Journal of Social Robotics*, 1–11. Online first publication.

Colombetti, G. (2014). *The feeling body: Affective science meets the enactive mind*. Cambridge, MA: MIT Press.

Curren, R. (2016). Aristotelian versus virtue ethical character education. *Journal of Moral Education*, *45*(4), 516–526.

Danaher, J. & McArthur, N. (Eds.). (2017). *Robot sex. social and ethical implications*. Cambridge, MA: MIT Press.

Danielson, P. (1992). *Artificial morality: Virtuous robots for virtual games*. London: Routledge.

Darley, J. M. & Batson, C. D. (1973). ‘‘From Jerusalem to Jericho”: A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology*, *27*(1), 100–108.

De Brigard, F. (2014). Is memory for remembering? Recollection as a form of episodic hypothetical thinking. *Synthese*, *191*(2), 155–185.

De Preester, H. (2010). Postphenomenology, embodiment and technics: Don Ihde, Postphenomenology and Technoscience: The Peking University Lectures. State University of New York Press, Albany, 2009 and

Embodied Technics. Automatic Press/VIP, 2010. *Human Studies*, *33*(2-3), 339–345.

De Preester, H. (2011). Technology and the body: The (im)possibilities of re-embodiment. *Foundations of Science*, *16*, 119–137.

de Boer, B. (2019). *How scientific instruments speak: A hermeneutics of technological mediations in (neuro-)scientific practice*. Enschede: Twente University.

de Graaf, M. M. A. (2016). An ethical evaluation of human–robot relationships. *International Journal of Social Robotics*, *8*(4), 589–598.

Deng, B. (2015). Machine ethics: The robot's dilemma. *Nature*, *523*(7558), 24–26.

Dennett, D. C. (1991). *Consciousness explained*. Boston, MA: Little, Brown.

———. (1993). Book review of *The embodied mind: Cognitive science and human experience*. *The American Journal of Psychology*, *106*(1), 121–126.

———. (2003). *Freedom evolves*. New York: Viking.

———. (2004). Calling in the Cartesian loans. *Behavioral and Brain Sciences*, *27*(5), 661.

———. (2005). *Sweet dreams: Philosophical obstacles to a science of consciousness*. Cambridge, MA: MIT Press.

Dhawan, N., Kunik, M. E., Oldham, J. & Coverdale, J. (2010). Prevalence and treatment of narcissistic personality disorder in the community: A systematic review. *Comprehensive Psychiatry*, *51*(4), 333–339.

Di Paolo, E. A. (2009). Extended life. *Topoi*, *28*(1), 9–21.

———. (2015). Interactive time-travel: On the intersubjective retro-modulation of intentions. *Journal of Consciousness Studies*, *22*(1–2), 49–74.

Di Paolo, E. A., Buhrmann, T. & Barandiaran, X. E. (2017). *Sensorimotor life: An enactive proposal*. Oxford: Oxford University Press.

Di Paolo, E. A., Cuffari, E. C. & De Jaegher, H. (2018). *Linguistic bodies: The continuity between life and language*. Cambridge, MA: MIT Press.

Doris, J. M. (1998). Persons, situations and virtue ethics. *Noûs*, *32*(4), 504–530.

———. (2002). *Lack of character: Personality and moral behavior*. Cambridge: Cambridge University Press.

Doyen, S., Klein, O., Pichon, C.-L. & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, *7*(1), e29081.

Dresler, M., Shirer, W. R., Konrad, B. N., Müller, N. C. J., Wagner, I. C., Fernández, G., … Greicius, M. D. (2017). Mnemonic training reshapes brain networks to support superior memory. *Neuron*, *93*(5), 1227–1235.

Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.

Engel, A. K. (2010). Directive minds: How dynamics shapes cognition. In J. Stewart, O. Gapenne & E. A. Di Paolo (Eds.), *Enaction: Toward a new paradigm in cognitive science* (pp. 219–243). Cambridge, MA: MIT Press.

Engel, A. K., Maye, A., Kurthen, M. & König, P. (2013). Where's the action? The pragmatic turn in cognitive science. *Trends in Cognitive Sciences*, *17*(5), 202–209.

Fassbender, E. & Heiden, W. (2006). The virtual memory palace. *Journal of Computational Information Systems*, *2*(1), 457–464.

Fernandes, K., Cardoso, J. S. & Astrup, B. S. (2018). A deep learning approach for the forensic evaluation of sexual assault. *Pattern Analysis and Applications*, *21*(3), 629–640.

Fernández, J. (2018). The functional character of memory. In K. Michaelian, D. Debus & D. Perrin (Eds.), *New directions in philosophy of memory* (pp. 52–72). New York: Routledge.

Fink, P. W., Foo, P. S. & Warren, W. H. (2009). Catching fly balls in virtual reality: A critical test of the outfielder problem. *Journal of Vision*, *9*(13), 14.

Fischer, J. M. & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge: Cambridge University Press.

Floridi, L. & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, *14*(3), 349–379.

Fodor, J. (1981). The mind-body problem. *Scientific American*, *244*(1), 114–123.

Fröding, B. E. E. (2011). Cognitive enhancement, virtue ethics and the good life. *Neuroethics*, *4*(3), 223–234.

Froese, T. (2014). Bio-machine hybrid technology: A theoretical assessment and some suggestions for improved future design. *Philosophy & Technology*, *27*(4), 539–560.

Gallagher, S. (2015). How embodied cognition is being disembodied. *The Philosophers' Magazine*, *68*, 96–102.

———. (2017). *Enactivist interventions: Rethinking the mind*. Oxford: Oxford University Press.

———. (2018). The extended mind: State of the question. *The Southern Journal of Philosophy*, *56*(4), 421–447.

Gallagher, S. & Zahavi, D. (2007). *The phenomenological mind: An introduction to philosophy of mind and cognitive science*. London: Routledge.

Gerrans, P. & Kennett, J. (2010). Neurosentimentalism and moral agency. *Mind*, *119*(475), 585–614.

Gibson, J. J. (1979). *The ecological approach to visual perception*. New York: Psychology Press.

Ginsborg, J. & Sloboda, J. A. (2007). Singers' recall for the words and melody of a new, unaccompanied song. *Psychology of Music*, *35*(3), 421–440.

Gips, J. (1995). Towards the ethical robot. In K. M. Ford (Ed.), *Android epistemology* (pp. 243–252). Cambridge, MA: MIT Press.

Godden, D. R. & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, *66*(3), 325–331.

Godfrey-Smith, P. (2016). *Other minds: The octopus, the sea, and the deep origins of consciousness*. Farrar, Straus and Giroux.

Goetz, J. L., Keltner, D. & Simon-Thomas, E. (2010). Compassion: An evolutionary analysis and empirical review. *Psychological Bulletin*, *136*(3), 351–374.

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist*, *54*(7), 493–503.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adversial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 27, pp. 2672–2680).

Güçlütürk, Y., Güçlü, U., Baró, X., Escalante, H. J., Guyon, I., Escalera, S., … van Lier, R. (2017). Multimodal first impression analysis with deep residual networks. *IEEE Transactions on Affective Computing*, 1–1.

Haney, C., Banks, C. & Zimbardo, P. G. (1973). Interpersonal dynamics of a simulated prison. *International Journal of Criminology and Penology*, *1*, 69–97.

Harman, G. (1988). Wide functionalism. In S. Schiffer & D. Steele (Eds.), *Cognition and representation* (pp. 11–20). Boulder, CO: Westview Press.

———. (1999). Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error. *Proceedings of the Aristotelian Society*, *99*(1), 316–331.

Harris, K. (2019). Whose (extended) mind is it, anyway? *Erkenntnis*, 1–15. Online first publication.

Haselager, P. (2013). Did I do that? Brain-computer interfacing and the sense of agency. *Minds and Machines*, *23*(3), 405–418.

Heersmink, R. (2012). Defending extension theory: A response to Kiran and Verbeek. *Philosophy & Technology*, *25*(1), 121–128.

———. (2017). Distributed selves: Personal identity and extended memory systems. *Synthese*, *194*(8), 3135–3151.

———. (2018). The narrative self, distributed memory, and evocative objects. *Philosophical Studies*, *175*(8), 1829–1849.

Hodges, S. D. & Wegner, D. M. (1997). Automatic and controlled empathy. In W. Ickes (Ed.), *Empathic accuracy* (pp. 311–339). New York: Guilford Press.

Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.

Howell, R. J. (2014). Google Morals, virtue, and the asymmetry of deference. *Noûs*, *48*(3), 389–415.

———. (2016). Extended virtues and the boundaries of persons. *Journal of the American Philosophical Association*, *2*(1), 146–163.

Humphreys, T. P. (2004). Understanding sexual consent: An empirical investigation of the normative script for young heterosexual adults. In M. Cowling & P. Reynolds (Eds.), *Making sense of sexual consent*. Ashgate.

Hutchins, E. (2005). Material anchors for conceptual blends. *Journal of Pragmatics*, *37*(10), 1555–1577.

Huttner, J.-P. & Robra-Bissantz, S. (2016). A design science approach to high immersive mnemonic e-learning. In *MCIS 2016 proceedings*. 28.

Hutto, D. D., Kirchhoff, M. D. & Myin, E. (2014). Extensive enactivism: Why keep it all in? *Frontiers in Human Neuroscience*, *8*(706).

Hutto, D. D. & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. Cambridge, MA: MIT Press.

Hutto, D. D. & Myin, E. (2017). *Evolving enactivism: Basic minds meet content*. Cambridge, MA: MIT Press.

Hutto, D. D. & Peeters, A. (2018a). The roots of remembering: Radically enactive recollection. In K. Michaelian, D. Debus & D. Perrin (Eds.), *New directions in philosophy of memory* (pp. 97–118). New York: Routledge.

———. (2018b). The roots of remembering: Radically enactive recollection. In K. Michaelian, D. Debus & D. Perrin (Eds.), *New Directions in Philosophy of Memory* (pp. 97–118). New York: Routledge.

Hutto, D. D., Peeters, A. & Segundo-Ortin, M. (2017). Cognitive ontology in flux: The possibility of protean brains. *Philosophical Explorations*, *20*(2), 209–223.

Ihde, D. (1990). *Technology and the lifeworld: From Garden to Earth*. Bloomington: Indiana University Press.

———. (2002). *Bodies in technology*. Minneapolis: University of Minnesota Press.

———. (2009). *Postphenomenology and technoscience: The Peking University lectures*. Albany: SUNY Press.

Ihde, D. & Malafouris, L. (2019). *Homo faber* revisited: Postphenomenology and material engagement theory. *Philosophy & Technology*, *32*(2), 195–214.

Janssen, J. H., Tacken, P., de Vries, J., van den Broek, E. L., Westerink, J. H., Haselager, P. & IJsselsteijn, W. A. (2013). Machines outperform laypersons in recognizing emotions elicited by autobiographical recollection. *Human–Computer Interaction*, *28*(6), 479–517.

Jimenez, M. (2016). Aristotle on becoming virtuous by doing virtuous actions. *Phronesis*, *61*(1), 3–32.

Jund, T., Capobianco, A. & Larue, F. (2016). Impact of frame of reference on memorization in virtual environments. In *2016 IEEE 16th international conference on advanced learning technologies* (pp. 533–537).

Kiran, A. H. & Verbeek, P.-P. (2010). Trusting our selves to technology. *Knowledge, Technology & Policy*, *23*(3-4), 409–427.

Kirchhoff, M. D. (2012). Extended cognition and fixed properties: Steps to a third-wave version of extended cognition. *Phenomenology and the Cognitive Sciences*, *11*(2), 287–308.

Kirsh, D. (2013). Embodied cognition and the magical future of interaction design. *ACM Transactions on Computer-Human Interaction*, *20*(1), 1–30.

Kong, G., He, K. & Wei, K. (2017). Sensorimotor experience in virtual reality enhances sense of agency associated with an avatar. *Consciousness and Cognition*, *52*, 115–124.

Kriegel, U. (2019). The intentional structure of moods. *Philosophers' Imprint*, *19*(49), 1–19.

Kristjánsson, K. (2015). *Aristotelian character education*. London: Routledge.

Krokos, E., Plaisant, C. & Varshney, A. (2019). Virtual memory palaces: Immersion aids recall. *Virtual Reality*, *23*(1), 1–15.

Kyselo, M. (2014). The body social: An enactive approach to the self. *Frontiers in Psychology*, *5*(986), 1–16.

Legge, E. L. G., Madan, C. R., Ng, E. T. & Caplan, J. B. (2012). Building a memory palace in minutes: Equivalent memory performance using virtual versus conventional environments with the method of loci. *Acta Psychologica*, *141*(3), 380–390.

Leman, M. & Maes, P.-J. (2014). Music perception and embodied music cognition. In L. Shapiro (Ed.), *The Routledge handbook of embodied cognition* (pp. 81–89). Oxford: Routledge.

Lemmens, P. (2008). *Gedreven door techniek: De menselijke conditie en de biotechnologische revolutie*. Oisterwijk: Box Press.

Lenggenhager, B., Tadi, T., Metzinger, T. & Blanke, O. (2007). Video ergo sum: Manipulating bodily self-consciousness. *Science*, *317*(5841), 1096–1099.

Levy, D. (2007a). *Intimate relationships with artificial partners*. Unpublished doctoral dissertation. Maastricht: Maastricht University.

———. (2007b). *Love and sex with robots: The evolution of human–robot relationships*. New York: Harper-Perennial.

———. (2012). The ethics of robot prostitutes. In P. Lin, K. Abney & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 223–232). Cambridge, MA: MIT Press.

Levy, N. (2007a). *Neuroethics: Challenges for the 21st century*. Cambridge: Cambridge University Press.

Levy, N. (2007b). Rethinking neuroethics in the light of the extended mind thesis. *American Journal of Bioethics*, *7*(9), 3–11.

———. (2013). Addiction is not a brain disease (and it matters). *Frontiers in Psychiatry*, *4*, 1–7.

———. (2014). Addiction as a disorder of belief. *Biology & Philosophy*, *29*(3), 337–355.

Lewis, M. (2017). Addiction and the brain: Development, not disease. *Neuroethics*, *10*(1), 7–18.

———. (2018). Brain change in addiction as learning, not disease. *The New England Journal of Medicine*, *379*, 1551–1560.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, *8*(4), 529–539.

Libet, B., Gleason, C. A., Wright, E. W. & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): The unconscious initiation of a freely voluntary act. *Brain*, *106*(3), 623–642.

Lim, G. Y. & Roloff, M. E. (1999). Attributing sexual consent. *Journal of Applied Communication Research*, *27*(1), 1–23.

Loader, P. (2013). Is my memory an extended notebook? *Review of Philosophy and Psychology*, *4*(1), 167–184.

Mackay, W. E., Fayard, A.-L., Probert, L. & Médini, L. (1998). Reinventing the familiar: Exploring an augmented reality design space for air traffic control. In *Conference proceedings on human factors in computing systems 1998* (pp. 558–565). New York: ACM Press/Addison-Wesley.

Madan, C. R. (2014). Augmented memory: A survey of the approaches to remembering more. *Frontiers in Systems Neuroscience*, *8*, 30.

Madan, C. R. & Singhal, A. (2012). Motor imagery and higher-level cognition: Four hurdles before research can sprint forward. *Cognitive Processing*, *13*(3), 211–229.

Maguire, E. A., Valentine, E. R., Wilding, J. M. & Kapur, N. (2002). Routes to remembering: The brains behind superior memory. *Nature Neuroscience*, *6*(1), 90–95.

Maiese, M. (2018). Can the mind be embodied, enactive, affective, *and* extended? *Phenomenology and the Cognitive Sciences*, *17*(2), 343–361.

———. (2019). Embodiment, sociality, and the life shaping thesis. *Phenomenology and the Cognitive Sciences*, *18*(2), 353–374.

Mäkelä, M. & Löytönen, T. (2017). Rethinking materialities in higher education. *Art, Design & Communication in Higher Education*, *16*(2), 241–258.

Malafouris, L. (2013). *How things shape the mind: A theory of material engagement*. Cambridge, MA: MIT Press.

Maturana, H. & Varela, F. J. (1980). *Autopoiesis and cognition: The realization of the living*. Dordrecht: D. Reidel.

Menary, R. (2007a). *Cognitive integration: Mind and cognition unbounded*. Basingstoke: Palgrave MacMillan.

———. (2007b). Writing as thinking. *Language Science*, *29*(5), 621–632.

———. (Ed.). (2010a). *The extended mind*. Cambridge, MA: MIT Press.

———. (2010b). The holy grail of cognitivism: A response to Adams and Aizawa. *Phenomenology and the Cognitive Sciences*, *9*(4), 605–618.

———. (2016). Pragmatism and the pragmatic turn in cognitive science. In A. K. Engel, K. J. Friston & D. Kragic (Eds.), *The pragmatic turn: Toward action-oriented views in cognitive science* (pp. 215–234). Cambridge, MA: The MIT Press.

Merritt, M. W., Doris, J. M. & Harman, G. (2010). Character. In J. M. Doris (Ed.), *The moral psychology handbook* (pp. 355–401). Oxford: Oxford University Press.

Metzinger, T. K. (2018). Why is virtual reality interesting for philosophers? *Frontiers in Robotics and AI*, *5*(101), 1–19.

Michaelian, K. (2016). *Mental time travel: Episodic memory and our knowledge of the personal past*. Cambridge, MA: MIT Press.

Michaelian, K., Debus, D. & Perrin, D. (Eds.). (2018). *New directions in philosophy of memory*. New York: Routledge.

Michaelian, K. & Sant'Anna, A. (2019). Memory without content? Radical enactivism and (post)causal theories of memory. *Synthese*. Online first publication.

Milgram, S. (1974). *Obedience to authority*. New York: Harper and Row.

Millikan, R. G. (1999). Historical kinds and the "special sciences". *Philosophical Studies*, *95*(1–2), 45–65.

Miranda, J. A., Canabal, M. F., Portela García, M. & Lopez-Ongil, C. (2011). Embedded emotion recognition: Autonomous multimodal affective internet of things. In F. Palumbo, C. Pilato, L. Pulina & C. Sau (Eds.),

Proceedings of the cyber-physical systems workshop 2018 (Vol. 2208, pp. 22–29). Alghero, Italy.

Morsella, E. (2005). The function of phenomenal states: Supramodular interaction theory. *Psychological Review*, *112*(4), 1000–1021.

Muensterer, O. J., Lacher, M., Zoeller, C., Bronstein, M. & Kübler, J. (2014). Google Glass in pediatric surgery: An exploratory study. *International Journal of Surgery*, *12*(4), 281–289.

Myin, E. & Zahnoun, F. (2018). Reincarnating the identity theory. *Frontiers in Psychology*, *9*(3), 1–9.

Nisbett, R. E. & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231–253.

Noë, A. (2004). *Action in perception*. Cambridge, MA: MIT Press.

Nørskov, M. (2015). Revisiting Ihde's fourfold "technological relationships": Application and modification. *Philosophy & Technology*, *28*(2), 189–207.

Noten, M., Peeters, A., van Toor, D., Winkens, L. & Jäkel, L. (2013). Hersenen, gedrag en middelengebruik: Een literatuurstudie naar de relatie tussen middelengebruik en geweld in het kader van straftoemeting. *Expertise en Recht*, *6*(4), 122–129.

Núñez, R., Allen, M., Gao, R., Miller Rigoli, C., Relaford-Doyle, J. & Semenuks, A. (2019). What happened to cognitive science? *Nature Human Behaviour*, 1–10.

Nussbaum, M. C. (2001). *The fragility of goodness: Luck and ethics in Greek tragedy and philosophy* (revised). Cambridge: Cambridge University Press.

Olson, E. T. (2011). The extended self. *Minds and Machines*, *21*(4), 481–495.

Olthof, B., Peeters, A., Schelle, K. & Haselager, P. (2013). If you're smart, we'll make you smarter: Applying the reasoning behind the development of honours programmes to other forms of cognitive enhancement. In F. Lucivero & A. Vedder (Eds.), *Beyond Therapy v. Enhancement? Multidisciplinary analyses of a heated debate* (pp. 117–142). RoboLaw. Pisa: Pisa University Press.

O'Regan, J. K. & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24*(5), 939–973.

Pacherie, E. (2006). Towards a dynamic theory of intentions. In S. Pockett, W. P. Banks & S. Gallagher (Eds.), *Does consciousness cause behavior?* (pp. 145–167). Cambridge, MA: MIT Press.

———. (2008). The phenomenology of action: A conceptual framework. *Cognition*, *107*(1), 179–217.

Peeters, A. (2017). Alexandru Dragomir: The world we live in [Book review]. *Phenomenological Reviews*, *3*, 54.

———. (in preparation. -a). Enactivism as a philosophy of technology.

———. (in preparation. -b). Is enactivism compatible with multiple realizability?

———. (in preparation. -c). Virtues, robots, and the enactive self.

Peeters, A. & Haselager, P. (2019). Designing virtuous sex robots. *International Journal of Social Robotics*, 1–12. Online first publication.

Peeters, A. & Segundo-Ortin, M. (2019). Misplacing memories? An enactive approach to the virtual memory palace. *Consciousness and Cognition*, *76*, 102834.

Philpot, R., Liebst, L. S., Levine, M., Bernasco, W. & Lindegaard, M. R. (2019). Would I be helped? Cross-national CCTV footage shows that intervention is the norm in public conflicts. *American Psychologist*, 231–253. Online first publication.

Piccinini, G. (2008). Computation without representation. *Philosophical Studies*, *137*(2), 205–241.

———. (2010). The mind as neural software? Understanding functionalism, computationalism, and computational functionalism. *Philosophy and Phenomenological Research*, *81*(2), 269–311.

———. (2015). *Physical computation: A mechanistic account*. Oxford: Oxford University Press.

Pickard, H., Ahmed, S. H. & Foddy, B. (2015). Alternative models of addiction. *Frontiers in Psychiatry*, *6*, 1–2.

Polger, T. W. (2004). *Natural minds*. Cambridge, MA: MIT Press.

Polger, T. W. & Shapiro, L. A. (2016). *The multiple realization book*. Oxford: Oxford University Press.

Prinz, J. (2009). The normativity challenge: Cultural psychology provides the real threat to virtue ethics. *The Journal of Ethics*, *13*(2), 117–144.

Putnam, A. L. (2015). Mnemonics in education: Current research and applications. *Translational Issues in Psychological Science*, *1*(2), 130–139.

Putnam, H. (1960). Minds and machines. In S. Hook (Ed.), *Dimensions of mind: A symposium* (pp. 138–164). New York: New York University Press. (Reprinted in Putnam [1975, pp. 362–385].)

———. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, mind, and religion* (pp. 37–48). Pittsburgh: University of Pittsburgh Press. (Reprinted as "The nature of mental states" in Putnam [1975, pp. 429–440].)

———. (1973). *Philosophy and our mental life*. Foerster Symposium on Computers and the Mind. University of California, Berkeley. (Reprinted in Putnam [1975, pp. 291–303].)

———. (1975). *Mind, language and reality*. Philosophical Papers. Cambridge: Cambridge University Press.

Ragan, E. D., Sowndararajan, A., Kopper, R. & Bowman, D. A. (2010). The effects of higher levels of immersion on procedure memorization performance and implications for educational virtual environments. *Presence: Teleoperators and Virtual Environments*, *19*(6), 527–543.

Raja, V. & Calvo, P. (2017). Augmented reality: An ecological blend. *Cognitive Systems Research*, *42*, 58–72.

Ramírez-Vizcaya, S. & Froese, T. (2019). The enactive approach to habits: New concepts for the cognitive science of bad habits and addiction. *Frontiers in Psychology*, *10*(301), 1–12.

Ransom, T. (2019). Process, habit, and flow: A phenomenological approach to material agency. *Phenomenology and the Cognitive Sciences*, *18*(1), 19–37.

Richardson, K. (2016). Sex robot matters: Slavery, the prostituted, and the rights of machines. *IEEE Technology and Society Magazine*, *35*(2), 46–53.

Rituerto-González, E., Mínguez-Sánchez, A., Gallardo-Antolín, A. & Peláez-Moreno, C. (2019). Data augmentation for speaker identification under stress conditions to combat gender-based violence. *Applied Sciences*, *9*(11), 2298.

Rosello, O., Exposito, M. & Maes, P. (2016). Nevermind: Using augmented reality for memorization. In *Proceedings of the 29th annual symposium on user interface software and technology* (pp. 215–216). ACM Press.

Rosenberger, R. & Verbeek, P.-P. (Eds.). (2015). *Postphenomenological investigations: Essays on human-technology relations*. Lanham: Lexington Books.

Rowland, I. D. (2008). *Giordano Bruno: Philosopher/heretic*. Chicago: University of Chicago Press.

Rupert, R. (2004). Challenges to the hypothesis of extended cognition. *Journal of Philosophy*, *101*(8), 389–428.

Sanderse, W. (2012). *Character education: A neo-Aristotelian approach to the philosophy, psychology and education of virtue*. Delft: Eburon.

———. (2015). An Aristotelian model of moral development. *Journal of Philosophy of Education*, *49*(3), 382–398.

Santoni de Sio, F. & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, *5*, 1–14.

Scheutz, M. (2012). The inherent dangers of unidirectional emotional bonds between humans and social robots. In P. Lin, K. Abney & G. A. Bekey (Eds.), *Robot ethics: The ethical and social implications of robotics* (pp. 205–222). Cambridge, MA: MIT Press.

Scheutz, M. & Arnold, T. (2017). Intimacy, bonding, and sex robots: Examining empirical results and exploring ethical ramifications. In J. Danaher & N. McArthur (Eds.), *Robot sex. social and ethical implications* (pp. 247–260). Cambridge, MA: MIT Press.

Schliesser, E. (2019). Synthetic philosophy. *Biology & Philosophy*, *34*(2), 1–9.

Senft, E., Lemaignan, S., Baxter, P. E., Bartlett, M. & Belpaeme, T. (2019). Teaching robots social autonomy from in situ human guidance. *Science Robotics*, *4*(35), eaat1186.

Shapiro, L. A. (2004). *The mind incarnate*. Cambridge, MA: MIT Press.

———. (2008). How to test for multiple realization. *Philosophy of Science*, *75*(5), 514–525.

———. (2019). Flesh matters: The body in cognition. *Mind & Language*, *34*(1), 3–20.

Sharkey, N. (2008). The ethical frontiers of robotics. *Science*, *322*(5909), 1800–1801.

Sharkey, N., van Wynsberghe, A., Robbins, S. & Hancock, E. (Eds.). (2017). *Our sexual future with robots: A Foundation for Responsible Robotics consultation report*. Foundation for Responsible Robotics.

Shirado, H. & Christakis, N. A. (2017). Locally noisy autonomous agents improve global human coordination in network experiments. *Nature*, *545*, 370–374.

Skorburg, J. A. (2019). Where are virtues? *Philosophical Studies*, *176*(9), 2331–2349.

Slors, M. (2015). Conscious intending as self-programming. *Philosophical Psychology*, *28*(1), 94–113.

———. (2019). Two distinctions that help to chart the interplay between conscious and unconscious volition. *Frontiers in Psychology*, *10*(552), 1–12.

Slote, M. (2001). *Morals from motives*. Oxford: Oxford University Press.

———. (2010). *Moral sentimentalism*. Oxford: Oxford University Press.

Smith, S. M. & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin & Review*, *8*(2), 203–220.

Soon, C. S., Brass, M., Heinze, H.-J. & Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, *11*(5), 543–545.

Sparrow, R. (2002). The march of the robot dogs. *Ethics and Information Technology*, *4*(4), 305–318.

———. (2016). Kicking a robot dog. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (p. 229). IEEE.

———. (2017). Robots, rape, and representation. *International Journal of Social Robotics*, *9*(4), 465–477.

———. (2019). Virtue and vice in our relationships with robots: Is there an asymmetry and how might it be explained? *International Journal of Social Robotics*. Manuscript under review.

Stapleton, M. (2013). Steps to a "Properly Embodied" cognitive science. *Cognitive Systems Research*, *22–23*, 1–11.

Sterelny, K. (2010). Minds: Extended or scaffolded? *Phenomenology and the Cognitive Sciences*, *9*(4), 465–481.

Stevens, C., Malloch, S., McKechnie, S. & Steven, N. (2003). Choreographic cognition: The time-course and phenomenology of creating a dance. *Pragmatics & Cognition*, *11*(2), 297–326.

Stoffregen, T. A., Bardy, B. G. & Mantel, B. (2006). Affordances in the design of enactive systems. *Virtual Reality*, *10*(1), 4–10.

Strikwerda, L. (2017). Legal and moral implications of child sex robots. In J. Danaher & N. McArthur (Eds.), *Robot sex. social and ethical implications* (pp. 133–152). Cambridge, MA: MIT Press.

Sutton, J. (2007). Spongy brains and material memories. In M. Floyd-Wilson & G. Sullivan (Eds.), *Embodiment and environment in early modern Europe* (pp. 14–34). London: Palgrave.

———. (2010). Exograms and interdisciplinarity: History, the extended mind, and the civilizing process. In R. Menary (Ed.), *The extended mind* (pp. 189–225). MIT Press.

Sutton, J. & Williamson, K. (2014). Embodied remembering. In L. A. Shapiro (Ed.), *The Routledge handbook of embodied cognition* (pp. 315–325). London: Routledge.

Swanton, C. (2003). *Virtue ethics: A pluralistic view*. Oxford: Oxford University Press.

Thagard, P. (2005). *Mind: Introduction to cognitive science* (2nd ed.). Cambridge, MA: MIT Press.

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind*. Cambridge, MA: Harvard University Press.

Thompson, E. & Stapleton, M. (2009). Making sense of sense-making: Reflections on enactive and extended mind theories. *Topoi*, *28*(1), 23–30.

Tonkens, R. (2012). Out of character: On the creation of virtuous machines. *Ethics and Information Technology*, *14*(2), 137–149.

Traeger, M., Sebo, S., Jung, M., Scassellati, B. & Christakis, N. (2019). *Vulnerable robots positively shape human conversational dynamics in a human-robot team*. Unpublished manuscript. Presented at Center for Empirical Research on Stratification and Inequality Spring 2019 Workshop at Yale University on January 31, 2019.

Vaccari, A. P. (2017). Against cognitive artifacts: Extended cognition and the problem of defining 'artifact'. *Phenomenology and the Cognitive Sciences*, *16*(5), 879–892.

Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford: Oxford University Press.

Van Waes, L. & Schellens, P. J. (2003). Writing profiles: The effect of the writing mode on pausing and revision patterns of experienced writers. *Journal of Pragmatics*, *35*(6), 829–853.

van Alphen, F. (2014). Tango and enactivism: First steps in exploring the dynamics and experience of interaction. *Integrative Psychological and Behavioral Science*, *48*(3), 322–331.

van Fraassen, B. C. (1980). *The scientific image*. Oxford: Oxford University Press.

Varela, F. J. (1993). Book review of *Consciousness explained*. *The American Journal of Psychology*, *106*(1), 126–129.

———. (1999). *Ethical know-how: Action, wisdom, and cognition*. Stanford: Stanford University Press.

Varela, F. J., Thompson, E. & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. Cambridge, MA: MIT Press.

Verbaarschot, C., Farquhar, J. & Haselager, P. (2015). Lost in time...: The search for intentions and readiness potentials. *Consciousness and Cognition*, *33*, 300–315.

Verbeek, P.-P. (2005). *What things do: Philosophical reflections on technology, agency, and design*. University Park, PA: Pennsylvania State University Press.

———. (2008). Cyborg intentionality: Rethinking the phenomenology of human–technology relations. *Phenomenology and the Cognitive Sciences*, *7*(3), 387–395.

———. (2011). *Moralizing technology: Understanding and designing the morality of things*. Chicago, IL: University of Chicago Press.

Wagman, J. B. & Chemero, A. (2014). The end of the debate over extended cognition. In T. Solymosi & J. Shook (Eds.), *Neuroscience, neurophilosophy and pragmatism* (pp. 105–124). Palgrave Macmillan.

Wallach, W. & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press.

Ward, D., Silverman, D. & Villalobos, M. (2017). Introduction: The varieties of enactivism. *Topoi*, *36*(3), 365–375.

Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, MA: MIT Press.

Wegner, D. M. & Bargh, J. A. (1998). Control and automaticity in social life. In D. T. Gilbert, S. T. Fiske & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 446–496). New York: McGraw-Hill.

Wegner, D. M., Erber, R. & Raymond, P. (1991). Transactive memory in close relationships. *Journal of Personality and Social Psychology*, *61*, 923–929.

Wheeler, M. (2010a). In defense of extended functionalism. In R. Menary (Ed.), *The extended mind* (pp. 245–270). MIT Press.

———. (2010b). Minds, things and materiality. In L. Malafouris & C. Renfrew (Eds.), *The cognitive life of things: Recasting the boundaries of the mind* (pp. 29–37). McDonald Institute monographs. Cambridge: McDonald Institute for Archaeological Research.

———. (2015). Extended consciousness: An interim report. *The Southern Journal of Philosophy*, *53*, 155–175.

———. (2017). The revolution will not be optimised: Radical enactivism, extended functionalism and the extensive mind. *Topoi*, *36*(3), 457–472.

Wilson, R. A. (1994). Wide computationalism. *Mind*, *103*(411), 351–372.

———. (2004). *Boundaries of the mind: The individual in the fragile sciences*. Cambridge: Cambridge University Press.

———. (2017). Group-level cognizing, collaborative remembering, and individuals. In M. L. Meade, P. Van Bergen, C. B. Harris, J. Sutton & A. J. Barnier (Eds.), *Collaborative remembering: Theories, research, and applications* (pp. 248–260).

Wilson, R. A. & Lenart, B. A. (2015). Extended mind and identity. In J. Clausen & N. Levy (Eds.), *Handbook of neuroethics* (pp. 423–439). Dordrecht: Springer.

Winfield, A. F. T., Blum, C. & Liu, W. (2014). Towards an ethical robot: Internal models, consequences and ethical action selection. In M. Mistry, A. Leonardis, M. Witkowski & C. Melhuish (Eds.), *Advances in autonomous robotics systems* (Vol. 8717, pp. 85–96). Lecture Notes in Computer Science. Springer.

Yamaji, Y., Miyake, T., Yoshiike, Y., De Silva, P. R. S. & Okada, M. (2011). STB: Child-dependent sociable trash box. *International Journal of Social Robotics*, *3*(4), 359–370.

Yates, F. A. (1966). *The art of memory* (ARK 1984 reprint). London: Routledge.

Zimbardo, P. G. (2007). *The Lucifer effect: Understanding how good people turn evil*. Oxford: Blackwell.

Publications

Peeters, **A.** (in preparation-a). Enactivism as a philosophy of technology.

Peeters, **A.** (in preparation-b). Free will put to the test: A compatibilist operationalization. *Philosophical Psychology*.

Peeters, **A.** (in preparation-c). Is enactivism compatible with multiple realizability?

Peeters, **A.** (in preparation-d). Virtues, robots, and the enactive self.

Cappuccio, M. L., **Peeters**, **A.** & MacDonald, W. D. (2019). Sympathy for Dolores: Moral consideration for robots based on virtue and recognition. *Philosophy & Technology*, 1–23. Online first publication.

Peeters, **A.** (2019). Steering away from multiple realization. *Adaptive Behavior*, 1–2. Online first publication.

Peeters, **A.** & Haselager, P. (2019). Designing virtuous sex robots. *International Journal of Social Robotics*, 1–12. Online first publication.

Peeters, **A.** & Segundo-Ortin, M. (2019). Misplacing memories? An enactive approach to the virtual memory palace. *Consciousness and Cognition*, *76*, 102834.

Hutto, D. D., Myin, E., **Peeters**, **A.** & Zahnoun, F. (2018). The cognitive basis of computation: Putting computation in its place. In M. Sprevak & M. Columbo (Eds.), *Handbook of the Computational Mind* (pp. 272–282). London: Routledge.

Hutto, D. D. & **Peeters**, **A.** (2018). The roots of remembering: Radically enactive recollection. In K. Michaelian, D. Debus & D. Perrin (Eds.), *New Directions in Philosophy of Memory* (pp. 97–118). New York: Routledge.

Hutto, D. D., **Peeters**, **A.** & Segundo-Ortin, M. (2017). Cognitive ontology in flux: The possibility of protean brains. *Philosophical Explorations*, *20*(2), 209–223.

Peeters, **A.** (2017a). Alexandru Dragomir: The world we live in [Book review]. *Phenomenological Reviews*, *3*, 54.

Peeters, **A.** (2017b). Freedom regained: The possibility of free will [Book review]. *Philosophical Psychology*, *30*(5), 682–684.

Noten, M., **Peeters**, **A.**, van Toor, D., Winkens, L. & Jäkel, L. (2013). Hersenen, gedrag en middelengebruik: Een literatuurstudie naar de relatie tussen middelengebruik en geweld in het kader van straftoemeting. *Expertise en Recht*, *6*(4), 122–129.

Olthof, B., **Peeters**, **A.**, Schelle, K. & Haselager, P. (2013). If you're smart, we'll make you smarter: Applying the reasoning behind the development of honours programmes to other forms of cognitive enhancement. In F. Lucivero & A. Vedder (Eds.), *Beyond Therapy v. Enhancement? Multidisciplinary analyses of a heated debate* (pp. 117–142). RoboLaw. (First authorship shared by BO, AP and KS.) Pisa: Pisa University Press.

Acknowledgments

It would be against the spirit of this dissertation to claim any credit solely for myself, as this was definitely a project distributed across me and my wider environment. In principle that would mean that the blame for any lingering mistakes in this dissertation can be distributed as well, but I will claim complete responsibility for those. In any case, many people helped shape my education and the present dissertation. The following is my best attempt at acknowledging them all, but invariably some might have been forgotten. To those I offer my sincere apologies.

The first to be mentioned, and rightly so, is Patrick McGivern, my principal advisor on this project. Patrick, you joined this project at a relatively late stage but your support has felt like the lifeline I needed when I was lost alone at sea. Your unwavering patience and insightful advice has made the conclusion of this project possible. You have been a tremendous pillar of support and I will aspire to become as great a mentor to others as you've been to me. I am grateful.

Rob Wilson gave an inspiring talk on his critical examination of eugenics – past and ongoing – at Wollongong in 2018, showing a combination of excellent philosophical skill and passionate societal engagement. I therefore felt incredibly fortunate when he accepted to join this project as my secondary advisor. Rob, though your presence on the team has been all too brief, it has been invaluable. I wish I had had more opportunities to learn from you and hope we can make up for that in the future.

If virtue is to be understood as spread out over agent and environment, then my friends, colleagues and students at Wollongong most definitely share credit for helping me become a better philosopher. Nick, your piercing mind and positive vibes have been almost as inspiring as your willingness to support those in need. I hope to see you and Linnea on many future visits. Alan and Miguel Segundo Ortin, you were my first

proper office mates and I couldn't have wished for better ones. Indeed, I will likely blame you two on many occasions for setting the bar too high. Fortunately, Russell Meyer was there to average it out a bit. Let's stay in touch over the adventures of good old Sam Pepys. Vern, you taught me my first Australian slang and helped me feel more at home here. All of you, let's make sure we discuss the newest *Star Wars* when it comes out. Farid, I feel lucky for having you visit us for half a year. Not only did you introduce me properly to the local legend that is Dicey Riley's, but you also allowed me to speak my native tongue at a time that I was feeling a bit homesick. Liz and Ding, I'm glad to have been there when you entered the field of philosophy. You're great philosophers and I hope we are able to have many more discussions on the things that matter most. Cameron, David, Ian, Jane, Jarrah, Keith, and Naomi, thank you all for having made my stay in Wollongong better.

Special mention should be made of the writing group that I regularly attended during the first two years of my doctoral candidature. Thank you Brian, Anu, Cathy and all the others. Your advice on healthy has been invaluable and, alas, has also definitely been ignored at some stages of this project.

Part of the research in this dissertation was conducted at the University of Edinburgh, Scotland. Andy Clark sponsored my visit there and I'm immensely grateful for the time he took to discuss matters of functionalism, extended mind, predictive processing, and science-fiction with me. You have been, and are, an inspiration for much of the research in the present text, even, or perhaps especially, on those topics that I disagree with you. Thanks also to John Dorsch, Mog Stapleton, Till Vierkant, Mike Wheeler and Lee Wilson for the discussions we had: I learned much from all of you. A special mention is reserved for Gavin, who turned out to be the best roommate I ever had. I hope I will one day finally manage to go to one of your gigs.

I want to thank the Hatestorm, X'Nedra, Mathilde, and Thudlan for the mercenary work we did together. Let's make sure we get the Tidings of Woe together again and venture forth from the city of Gloomhaven. You know who you are.

My gratitude also extends to those who saw me leave to the other side of the world. Some of you actually came to visit, but even if you didn't,

your continued support helped me get this far. Thank you, Jasper, Joyce, Bas L., Judith, Maarten, Bram, Suze, Saskia, Bas F., Maaike and Frank.

It has been a long time since I first started studying philosophy. I've shared many experiences with my friends from the 'reading group' and am grateful for their continued support and shared love for philosophy. Let's drink some cassis again soon, Jitse, Paulien, Jonne, Rob, and especially Jorrit.

Thanks also to my previous mentors and teachers, who fanned my passion for philosophy: Marc Slors, Pim Haselager, Pieter Lemmens, and Ad Vennix. And to Jos Kusters, who helped spark the initial flame and who once with mock-seriousness admonished me by saying that the topic of my *profielwerkstuk* was large enough to inspire a dissertation – little did I know what a dissertation was but the word stuck. I would not have been where I am if not for every single one of you.

There are some cases were it is hard to draw the line on what to include and who not. The bounds of my extended family have been fluid for most of my life. I cannot mention them all here, but I will mention a few. Thank you Joop and Janneke for coming to visit us in Australia every year and laughing heartily at me ducking away in the cinema. Jason and Michel, I love you my brothers. Zahi, you should've done that internship in Sydney. Then again, this thesis might've taken a lot longer then so maybe it's for the best. Let's take some trips in Germany together. Mom, dad, thanks for believing in me. And thanks again to you dad, for recommending that I should take a look into this thing called 'philosophy', even though you disliked it so yourself.

While I am writing these words, the best part of my enactive self is at the other side of the world. Without her, many of the illustrations in this dissertation would not have been possible – or would have looked a lot worse. We went on quite the adventure together and it looks like the next one is just around the corner. I don't think I can top what I said in my master's thesis, so I'll just say this: Lies, I look forward to adventuring with the three of us soon.

By way of opening the "Big Ideas Festival" held at the University of Wollongong on the 3rd of October 2017, Jade Kennedy of the Aboriginal Yuin people related a powerful story about hospitality. I couldn't possibly tell the story as well as he did, both because of his narrative talent and our

different backgrounds, but I will attempt to convey its main message. The story relates a man opening his home to a visitor. For each room, the man would say what is alright and what isn't. "This is the living room. Feel free to make yourself at home and relax on the couch. But please, do not put your feet on the table and if you like to smoke, do so in the yard." Kennedy then concluded his story by welcoming us at his home.

It has barely been two years and my current institution has decided, against a storm of protest, to put its feet on the table anyway and host a degree on Western Civilisation while having at the same time reduced the unit for Aboriginal Studies. The 'WestCiv' degree is advertised as being 'inclusive'. It is easy to appear magnanimous and initiate dialogue when the other party barely receives a stage to speak. I am appalled by this development and ashamed by some of my fellow philosophers' support of it, though I hasten to say that there have been great colleagues in the discipline who have spoken out against it. The research in the present dissertation has been done on the soil of the Dharawal Country of which I acknowledge the Wadi Wadi people as the traditional custodians. I stand with them in solidarity with struggles historical and ongoing against the oppression of Aboriginal Australians. Let's get our feet off the bloody table.

Curriculum vitæ

Anco Peeters was born to a life of unfulfilment on October 1, 1986 in 's-Hertogenbosch, the Netherlands. Thinking that studying philosophy would satiate his curiosity, he was sorely disappointed when he obtained his bachelor's degree in that area (2010) from Radboud University, Nijmegen, the Netherlands, and did not feel he knew more than when he started. After engaging with student representation in various capacities in local and national student unions, he stubbornly decided to continue his studies, receiving a bachelor's degree in Artificial Intelligence (2017) and a master's in Philosophy of Mind & Science (2015, *cum laude*) from the same university. Unfortunately, this made him realise he knew even less than he thought he did and he obtained scholarships to continue being disappointed and start his doctoral studies at the University of Wollongong, Australia. He gained some measure of – some would even say 'sadistic' – pleasure out of inflicting feelings of curiosity and lack of knowledge upon others, teaching subjects both to students of philosophy and computer science. As part of his doctoral studies, he was invited to visit the University of Edinburgh, Scotland from October to December 2018. Though having completed his doctoral studies in 2019, of which the present dissertation is the result, Anco's curiosity has still not been quenched and he has accepted a postdoctoral research position in philosophy of memory at the Ruhr-University Bochum, Germany. He is comforted by the recent realisation that, having read around somewhat, he might not be the only one who knows less than they think they do. More on his work can be found online, at www.ancopeeters.com.