Philos. Technol. (2018) 31:629–653
https://doi.org/10.1007/s13347-018-0317-3

RESEARCH ARTICLE

Toward an Ethics of AI Assistants: an Initial Framework

John Danaher[1]

Received: 22 October 2017 / Accepted: 22 May 2018 / Published online: 26 June 2018

Abstract Personal AI assistants are now nearly ubiquitous. Every leading smartphone operating system comes with a personal AI assistant that promises to help you with basic cognitive tasks: searching, planning, messaging, scheduling and so on. Usage of such devices is effectively a form of algorithmic outsourcing: getting a smart algorithm to do something on your behalf. Many have expressed concerns about this algorithmic outsourcing. They claim that it is dehumanising, leads to cognitive degeneration, and robs us of our freedom and autonomy. Some people have a more subtle view, arguing that it is problematic in those cases where its use may degrade important interpersonal virtues. In this article, I assess these objections to the use of AI assistants. I will argue that the ethics of their use is complex. There are no quick fixes or knockdown objections to the practice, but there are some legitimate concerns. By carefully analysing and evaluating the objections that have been lodged to date, we can begin to articulate an ethics of personal AI use that navigates those concerns. In the process, we can locate some paradoxes in our thinking about outsourcing and technological dependence, and we can think more clearly about what it means to live a good life in the age of smart machines.

Keywords Artificial intelligence · Degeneration · Cognitive outsourcing · Embodied cognition · Autonomy · Interpersonal communications

1 Introduction

Personal AI assistants are now almost ubiquitous. Every smartphone operating system comes with a personal AI assistant that promises to help you with basic cognitive tasks: searching, planning, messaging, scheduling and so on. Google's Assistant, Apple's Siri and Microsoft's Cortana are the obvious examples. But they are just the tip of a large

✉ John Danaher
johndanaher1984@gmail.com; john.danaher@nuigalway.ie

[1] School of Law, NUI Galway, University Road, Galway, Ireland

and ever-growing iceberg. Specialised apps offer similar assistance for specific tasks. Although these AI assistants are currently in their infancy, we can expect the underlying technology to develop and for the usage of AI assistants to expand.

This raises some interesting ethical and normative questions. As Selinger and Frischmann (2016) have recently noted, usage of AI assistance is effectively a new form of *outsourcing*. Humans have long outsourced the performance of cognitive tasks to others. I don't do my tax returns; my accountant does. I don't book my travel arrangements; my assistant does. Such humanistic outsourcing has its own ethical issues. If I get someone to do something on my behalf I need to ensure that they do so voluntarily, that they are fairly compensated, and that they are not exploited. On top of this, as Michael Sandel (2012) has argued, there are some tasks that seem to ethically demand my personal involvement. For instance, outsourcing the writing of a best man's speech seems like a mark of disrespect and apathy, not a praiseworthy efficiency-maximising way to fulfil one's duties.

If humanistic outsourcing demands its own ethical framework, then presumably AI outsourcing does too. But what might that ethical framework look like? There is no shortage of opinions on this matter. Some people think the answer is simple. They think that AI outsourcing is generally problematic. It is dehumanising (Kelly and Dreyfus, 2011; Frischmann 2014), leads to cognitive degeneration (Carr 2014) and robs us of our freedom and autonomy (Krakauer 2016; Crawford 2015). If it is to be done at all, it should be done sparingly and wisely. Some people have a more subtle view (Selinger 2014a, 2014b, 2014c), arguing that it is problematic in those cases where its use may degrade important interpersonal virtues.

In this article, I want to assess these objections to the use of AI assistants. Contrary to the view stated in the previous paragraph, I will argue that the ethics of AI outsourcing is complex in the sense that there are no quick fixes or knockdown objections to the practice, although there are some legitimate concerns. By carefully analysing and evaluating those concerns, we can begin to articulate an ethical framework for deciding when it is appropriate to make use of an AI assistant. In the process, we can locate some paradoxes in our thinking about outsourcing and technological dependence, and we can think more clearly about what it means to live a good life in the age of smart machines.

I proceed in six parts. In Section 2, I clarify the target for the remainder of the article and present a general theoretical model for thinking about the ethics of AI usage. I then proceed to examine various objections to the practice of AI outsourcing in descending order of generality. I start with the most common and general objection—the degeneration objection—in Section 3. I then look at some freedom/autonomy related objections in Section 4. In Section 5, I turn to a more discrete objection, though still one with a degree of generality, focusing on personal virtues and interpersonal relationships. In Section 6, I discuss the paradox of internal automaticity. And in Section 7, I conclude by distilling the lessons and paradoxes of the preceding analysis.

2 What are Artificially Intelligent Personal Assistants?

Let me start by clarifying the phenomenon of interest. This is not an easy task. Although the term 'artificial intelligence' is widely used, its precise definition is

contested. For example, Russell and Norvig, in their leading textbook on AI, discuss eight different definitions, divided into four main categories: thinking like a human, acting like a human, thinking rationally, and acting rationally (Russell and Norvig, 2016, 2 ff). Classically, following the work of Alan Turing, human-likeness was the operative standard in definitions of AI. A system could only be held to be intelligent if it could think or act like a human with respect to one or more tasks. Over time, rationality, conceived in instrumental/goal-directed forms, has become the more preferred standard. Indeed, Russell and Norvig themselves opt for this definition, describing as an 'AI' any system that 'acts so as to achieve the best outcome or, when there is uncertainty, the best expected outcome' (Russell and Norvig, 2016, 3). A similar definition was favoured by John McCarthy—often described as the father of AI—who described intelligence as the 'computational part of the ability to achieve goals in the world' and AI as the science and engineering of creating intelligent systems.[1]

In this paper, I follow this 'rationality' standard and will define as a personal AI assistant any computer-coded software system/program that can act in a goal-directed manner. In other words, I define it as a program that can be set some target output ('find me the best restaurant in my local area') and can select among a range of options that optimises (according to some specific metric) for that output. This definitional approach is generous, and includes broad and narrow forms of AI within its scope (i.e. AI capable of general problem-solving across multiple domains or AI capable of solving one or two problems in discrete domains). For the purposes of this article, I am concerned with the ethics of using such AI in our *personal lives*, not with the ethics of such AI in public (e.g. military or government) or commercial contexts. That is to say, I am interested in exploring the ethical dimension to the use of AI in personal decision-making, goal setting and goal achievement—the cases where you use an AI assistant to pick a movie to watch or a meal to cook, to plan your travel arrangements, to book appointments, to pay for services, to file taxes, to send messages, to perform household chores, and, more generally, to solve your personal problems.

This narrow focus is not intended to ignore or downplay the public and commercial uses of AI. Most of the technologies I talk about are commercial in nature, and individuals might use them in the furtherance of commercial or public aims (if those aims line up with their own personal goals). But I am ignoring the ethical questions that might arise from those public or commercial uses of the technology. I argue that this limitation of focus is both legitimate and appropriate. Distinctive ethical issues arise from the public and commercial use of AI (e.g. issues around democratic participation, privacy and public accountability, bias, procedural fairness and so on), and some of those issues have been analysed at considerable length already (e.g. Danaher 2016a; Mittelstadt et al. 2016). The ethics of AI assistance in the personal domain is relatively neglected by comparison.

AI assistance in the personal domain can be viewed as a form of *algorithmic cognitive outsourcing*, i.e. the offloading of a cognitive task to a smart algorithm. This makes AI assistance a species of automation. Automation is the general phenomenon, whereby once human (or animal) -performed tasks are performed by machines. Much historic automation has involved machines taking over the physical, non-cognitive elements of human or

[1] The quotes come from John McCarthy 'What is Artificial Intelligence? Basic Questions' available at http://www-formal.stanford.edu/jmc/whatisai/node1.html – note that this quote and the quote from Russell and Norvig was originally sourced through Scherer 2016

animal tasks. With the growth of AI, we see automation creeping into the mental and cognitive elements of tasks. In many ways, this is what makes the rise of AI assistance so ethically contentious. In modern societies, it is to the mental and cognitive that we attach much of our self-worth and social value. This forms an essential underlying assumption for the ethical evaluation undertaken in the remainder of this paper.

In order to think about the ethical significance of such cognitive outsourcing, it helps to draw upon the theoretical models proposed within the situated/embodied cognition literature (indeed, this literature is explicitly invoked in many of the arguments discussed below).[2] The central thesis of the situated/embodied cognition school of thought is that cognition is not a purely brain-based phenomenon (Kirsh 2010 & 1995; Norman 1991; Heersmink 2013 & 2015; Crawford 2015). We don't just think inside our heads. Our bodies and environments shape the way we perceive and process cognitive tasks. Cognition is a *distributed phenomenon* not a localised one, i.e. the performance of a cognitive task is something that gets distributed across brains, bodies and environments. I am not going to defend this situated/embodied view of cognition in this article. I think it is right in its basic outline, but there are legitimate criticisms of particular positions taken up within the literature. I appeal to it here because I think it provides a useful model for understanding both the phenomenon of AI assistance and its discontents.

One way that it helps us to do this is by identifying and explaining the long-standing importance of *cognitive artifacts* in human thinking. Cognitive artifacts can be defined as tools, objects or processes that assist in the performance of a cognitive task (Norman 1991; Heersmink 2013). Artifacts of this sort are abundant:

> We use maps to navigate, notebooks to remember, rulers to measure, calculators to calculate, sketchpads to design, agendas to plan, textbooks to learn, and so on. Without such artifacts we would not be the same cognitive agents, as they allow us to perform cognitive tasks we would otherwise not be able to perform.
>
> (Heersmink 2013, 465–466)

AI assistants can be viewed as simply a new type of cognitive artifact. The crucial question is whether they have unique and distinctive ethical consequences.

To answer that question, it helps to adopt two further ideas from the situated/embodied analysis of cognitive artifacts. The first is that although cognitive artifacts often enhance our performance of cognitive tasks—I am undoubtedly a better mathematician with a pen and paper than I am without—they do so in a particular way. We can think about our interactions with cognitive artifacts at the *system level* (i.e. our brains/bodies *plus* the artifact) and the *personal level* (i.e. how we interact with the artifact). The distinction comes from Norman (1991). At the system level, the cognitive performance is often enhanced by the artifact: me-plus-pen-and-paper is better at than me-without-pen-and-paper. But the system level enhancement is achieved by *changing the cognitive task* performed at the personal level: instead of imagining numbers in my

[2] Indeed, this literature is explicitly invoked by many of the critics of AI assistance e.g. Carr 2014, Krakauer 2016, and Crawford 2015.

head and adding and subtracting them using some mentally represented algorithm, I visually represent the numbers on a page, in a format that facilitates the easy application of an algorithm. The artifact changes one cognitive task into another (series) of cognitive tasks.

Why is this important? The answer (and the second key idea) is that it encourages us to think about the effects of cognitive artifacts in an *ecological* mode. So that when we start using a new artifact to assist with the performance of a cognitive task, we shouldn't think of this simply as a form of outsourcing. The artifact may share (or takeover) the cognitive burden, but in doing so, it will also change the cognitive environment in which we operate. It will create new cognitive tasks for us to perform and open up new modes or styles of cognition. For example, by performing computations with a pen and paper, we will be able to do far more complex mathematical operations than we could without. Changing one aspect of the cognitive environment has knock-on effects on other aspects of the cognitive environment.

This ecological model helps us to understand why one of the most popular dismissals of techno-pessimists (including critics of AI) is so attractive and so misleading. It is common, whenever someone expresses concerns about the cognitive consequences of a new technology like AI assistants, to bring up the Platonic dialogue *The Phaedrus* as a *reductio* of those concerns. In that dialogue, Socrates has a debate with Phaedrus about the merits of writing vis-a-vis oratory. Socrates insists that writing is bad for thinking. It weakens your innate cognitive capacities, makes you dependent on the artifact of the written word, and less able to remember and think complex thoughts for yourself. As he put it:

> …Their trust in writing, produced by external characters which are no part of themselves, will discourage the use of their own memory within them. You have invented an elixir not of memory, but of reminding; and you offer your pupils the appearance of wisdom, not true wisdom, for they will read many things without instruction and will therefore seem to know many things, when they are for the most part ignorant and hard to get along with, since they are not wise, but only appear wise.
>
> (Plato 1925 *The Phaedrus* 274d)

The view that Socrates expresses seems quaint and silly to modern ears. It seems as though he did not appreciate all the advantages that writing could bring. By writing down our thoughts and ideas, we allowed for their cultural transmission and preservation; we allowed for others to interpret, critique and build upon them. In short, we dramatically changed the cognitive ecology in which we live and breathe. These changes have enabled us to soar to new cognitive and civilisational heights.

And yet what Socrates said seems quite similar to what modern-day technological naysayers have to say about AI. They too worry about the cognitive consequences of the latest technological aids; and they too worry that these technological aids will make us stupid, more dependent and less able. So can we dismiss their concerns as easily as we dismiss Socrates's? I think not. *The Phaedrus* is not a *reductio* of the modern-day AI naysayers. There are two reasons for this. First, doing so fails to engage with the

particularities of their criticisms. The way in which writing changes our cognitive ecology is not necessarily the same as the way in which AI assistants change our cognitive ecology. Although they all belong to the general family of cognitive artifacts, the members of that family differ in their uses and their consequences. We need to attend to those differences when developing an appropriate ethics of AI assistance. Second, *The Phaedrus* is only persuasive with the benefit of hindsight. It is only because we see the advantages that writing has wrought that we are able to dismiss Socrates's view. We don't have the benefit of hindsight when it comes to AI assistants. We can see some of the changes they may bring, but not all. This uncertainty needs to be factored into our ethical analysis and it makes it less easy to dismiss the concerns of the critics.

Consequently, I think it is important to give the critics' views a fair hearing and to assess the merits of those criticisms in light of the features of our cognitive ecology that they call out. With this goal in mind, in the remainder of this article, I will assess three major objections to the use of personal AI assistants. Each of the objections has been defended by at least one sceptic of this technology. And they have been chosen for analysis because I believe them to be paradigmatic, though not exhaustive ,[3] of the kinds of concerns raised in both the academic literature and among the general public. Furthermore, each of the arguments identifies a set of ecological consequences arising from the use of an AI assistant, some of which are quite general, others more narrow. Assessing them will give good coverage of the sorts of effects that AI assistants might cause within our changing cognitive ecology.

3 The Degeneration Argument

With that said, I will to start by looking at the contemporary argument that is most similar to the one defended by Socrates in the *Phaedrus* and claims the most general ecological effect of the use of AI assistants. This is the *degeneration argument*. Though variants of this argument are commonplace, I will focus on the version presented in Nicholas Carr's book *The Glass Cage* (2014). In this book, he tries to marshal a number of philosophical and psychological ideas into a general critique of automation. The book focuses on all forms of automation, but it is clear that AI-based automation is the major concern. And although some of Carr's criticisms are concerned with the more social and political effects of automation, the degeneration argument is focused directly on the personal consequences of automation.

Carr defends the degeneration argument indirectly by first considering a contrary point of view: that of Alfred North Whitehead. In his 1911 work *An Introduction to Mathematics*, Whitehead made a bold claim:

[3] A reviewer wonders, for example, why I do not discuss the consequences of using AI assistants to outsource moral decision-making. There are several reasons for this. The most pertinent is that I have discussed moral outsourcing as a specific problem in another paper (Danaher 2016b) and, as I point out in that paper, I suspect discussions of moral outsourcing to AI will raise similar issues to those already discussed in the expansive literature on the use of enhancement technologies to improve moral decision-making (for a similar analysis, coupled with a defence of the use of AI moral assistance, see Giublini and Savulescu 2018). That said, some of what I say below about degeneration, autonomy and interpersonal virtue will be also be relevant to debates about the use of moral AI assistance.

> It is a profoundly erroneous truism, repeated by all copy-books and by eminent people when they are making speeches, that we should cultivate the habit of thinking of what we are doing. The precise opposite is the case. Civilization advances by extending the number of important operations which we can perform without thinking about them. Operations of thought are like cavalry charges in a battle — they are strictly limited in number, they require fresh horses, and must only be made at decisive moments.
>
> (Whitehead 1911, 45–46)

Though Whitehead was writing long before modern AI, what he says has relevance to the debate about personal AI assistants. He suggests that mental resources are limited and need to be saved—like the cavalry charges in a battle—for the times that really matter. Such savings are exactly what AI assistants promise. AI assistants can reduce the amount of thinking we need to do by automating certain cognitive tasks and thereby freeing up our mental resources for the 'decisive moments'.

So it seems like we can co-opt Whitehead's claims into the following defence of AI assistants (note: this argument is not intended to be formally valid):

(1) Mental labour is difficult and finite: time spent thinking about trivial matters limits our ability to think about more important ones.
(2) It is good when we have the time and ability to think the more important thoughts.
(3) Therefore, it would be good if we could reduce the amount of mental labour expended on trivial matters and increase the amount spent on important ones.
(4) AI assistants help to reduce the amount of mental labour expended on trivial matters.
(5) Therefore, it would be good if we could outsource more mental operations to AI assistants.

Although he doesn't present his analysis in these terms, Carr's defence of the degeneration argument starts by highlighting the flaws in Whitehead's inference from (3) and (4) to (5). The problem with this inference is that it assumes that if AI assistance saves mental labour, we will use the reserved mental labour to think more important thoughts. This ignores the potential knock-on effects of increased reliance on AI assistance. One of those knock-on effects, according to Carr, is that increased reliance on AI assistance will atrophy and degenerate our mental faculties. So, far from freeing up mental resources, increased reliance on AI assistance will deplete mental resources. We will no longer have the ability to think the important thoughts. This will in turn reduce the quality of our personal lives because the ability to engage in deep thinking is both intrinsically and instrumentally valuable: it results in a better immediate conscious experience and engagement with life, and it helps one to solve personal problems.

So Far, So Socrates. The novelty in Carr's argument comes from the psychological evidence he uses to support it. In the 1970s, psychologists discovered something they called the *generation effect* (Slamecka and Graf 1978). The original experiments in which it was discovered had to do with memorisation and recall. The finding was that the more cognitive work you have to do during the memorisation phase, the better able you are to recall the information at a future date. Later studies revealed that this effect

applied outside of memorisation: it helped with conceptual understanding, problem solving, and recall of more complex materials too.

The generation effect has a corollary: *the degeneration effect*. If anything that forces us to use our own internal cognitive resources enhances our memory and understanding, then anything that takes away the need to exert those internal resources will reduce our memory and understanding. Carr cites the experimental work of Christof van Nimegen and his colleagues in support of this view (Van Nimwegen et al. 2006; Burgos et al. 2007). They have worked on the role of assistive software in conceptual problem-solving. In one of their studies (van Nimwegen et al. 2006), they presented experimental subjects with a variation on the Missionaries and Cannibals game (a classic logic puzzle about getting a group of missionaries across a river without being eaten by a cannibal). The game comes with a basic set of rules. You must get the missionaries across the river in the least number of trips while conforming to those rules. Van Nimwegen and his colleagues got one group of subjects to solve the puzzle with a simple software program that provided no assistance and a second group to solve it using a software program that offered on-screen prompts, including details as to which moves were permissible. People using the assistive software solved the puzzles more quickly than the others. But in the long-run, the second group emerged as the winners: they solved the puzzles more efficiently and with fewer wrong-moves. What's more, in a follow-up study performed 8 months later, it was found that members of the second group were better able to recall how to solve the puzzle. Subsequent studies (Burgos et al. 2007) showed that the effect was also found for other cognitive tasks.

This adds a degree of evidential credibility to the Socratic-style dependency objection. But does it succeed in defeating Whitehead's argument? Not necessarily. Remember, we need to think about the effects of AI assistance in an ecological mode. It is not enough to prove that relying on AI assistance for certain cognitive tasks will lead to the degeneration in our ability to perform those tasks. I am willing to grant that degeneration is a plausible consequence of reliance on AI: everything we know from the psychology and neuroscience of human performance suggests that the brain has some plasticity: if you stop performing certain tasks the relevant cognitive real estate can be redeveloped and used for other tasks. But degeneration of performance on certain tasks is not enough, in and of itself, to show that there is a problem. It could well be that tasks in which our performance degenerates are not that important in the first place, and freeing us from their performance might be a net benefit. Indeed, this is often provably the case at an individual level. For example, I am no good at tracking, recording, and budgeting my finances. I find the task mind-numbingly boring. I have tried to do it several times in the past and failed. Fortunately, my bank recently started offering a software service that automatically categorises my expenditure items (this was easy since I conduct nearly all transactions electronically), presented nice visual readouts of this information, and gave me budgeting and saving suggestions. This has had a dramatic and immediate effect on my behaviour. I am now more aware of what I am spending, and saving more money per month. It is a clear and simple objective metric of success. I have no doubt that this automated expenditure tracking and budgeting system is causing some cognitive degeneration (I no longer even try to keep a mental track of my expenditure on a daily basis), but this degeneration is having net benefits. My overall cognitive ecology has changed for the better. I am better at solving the financial problems that life throws in my way.

To claim that degeneration is a general reason to object to the use of AI assistants, Carr will have to do more than prove the likelihood of degeneration for some specific task. He will have to show that the degeneration effect is either non-localised or that it is likely to affect some specific ability or set of abilities that is so intrinsically valuable that its degeneration would make life much worse. I consider the latter possibility in subsequent sections, particularly in Section 5 when I look at the possible effects of AI assistants on interpersonal relationships. For now, I focus on the general case, since it seems to be more in keeping with Carr's aspirations.

To succeed in making the general case, Carr will have to show that reliance on AI (i) degenerates cognition beyond some individual task and/or (ii) that the rest of our cognitive ecology will be changed so dramatically that it compounds the degeneration effect in the localised domain. To be fair, Carr has tried to do this. He himself (2011) along with several other authors (Crawford 2015; Newport 2016) have been arguing for some time now that ICT has dramatically changed our cognitive environment to the point that we are incapable of serious cognitive effort. The Internet, they say, is giving birth to a generation of information junkies, addicted to a constantly changing, constantly updating soup of social media posts, fake news and cat videos. People who live in such cognitive environments are never going to think the big thoughts that Whitehead revered. What's more, there is evidence to suggest that this distraction-rich environment has non-localised effects on our cognition. Clifford Nass's research, for example, suggests that people who constantly live and work in a distraction-rich environment are less able to focus on important tasks when the need arises (Nass and Flatow 2013 and Ophir et al., 2009). They are, in his own words, 'chronically distracted' and 'suckers for irrelevancy' (Nass and Flatow 2013). Incorporating this insight makes the degeneration argument more interesting and more challenging. The claim now is that the degenerating effect of AI comes on top of the pre-existing degeneration effect of our distraction-rich cognitive ecology. The result is that the degeneration effect spreads beyond the particular forms of assistance that AI provides. If we continue down the path to more AI assistance, we risk losing the ability to think deep thoughts by ourselves.

But even this does not provide a knockdown objection to the personal use of AI. All it does is force a nuanced and careful approach to the ethics of AI outsourcing. There are three reasons for this.

First, the value assumption underlying this general version of the degeneration argument must be kept in mind. The assumption is that the capacity to think deep thoughts is both intrinsically and instrumentally valuable. There is credibility to this. Understanding something, or working out the solution to a problem, can be intrinsically rewarding. There is a degree of conscious satisfaction/achievement that comes from the personal achievement that is lost if you outsource a task to another person or thing. Also, solving a cognitive problems is clearly instrumentally rewarding: it enables us to get other good things (food, money, shelter, success, status etc.) that make life worth living. Indeed, according to one theory (Pinker 2010), the ability to solve cognitive problems is fundamental to why humans still exist to this day. But the intrinsic and instrumental rewards of thinking are dissociable, and this has an important impact on the ethics of AI outsourcing. If the primary good of cognition is in the instrumental rewards it brings, then it is not clear that we lose anything by excessive use of AI assistants. Indeed, we may gain more than we lose. Take, once again, my example of the expenditure tracking and budgeting advice that is now automatically given out by my bank. This is

instrumentally valuable to me. I am saving more money as a result of this AI assistance, and this enables me to get more of what I want in the future. Furthermore, since I did not enjoy the mental effort involved in solving this cognitive problem (it was not intrinsically rewarding to me), it seems like this particular instance of using an AI assistant is a win-win, despite the degenerating effects it may have. The same logic can apply across a whole suite of AI assistants and their degenerating effects. As long as I am embedded in a network of devices that brings me the same (or better) instrumental rewards, and as long as I only rely on those devices that solve problems I did not enjoy solving, there is little to lament. The only problem from an instrumental perspective is if I become *decoupled* from the network of devices, or if the network of devices *breaks down*. In those cases, there will be knock-on negative effects to my well-being. Building up a degree of independent cognitive resilience could save me from these negative effects, but should I want to do this? That really depends on how likely decoupling and breakdown are, and how severe the effects will be. The reality is that in the modern era total disconnection is relatively rare and so you may waste mental effort by trying to build up cognitive resilience. Furthermore, as Heersmink argues (2015), the existence of these networks changes the instrumental rewards that are available to us. It could be that I am rewarded for being an expert user and navigator of my AI-rich ecology. If so, voluntarily decoupling myself from the system could bring more costs than rewards.

This brings me to the second point. One thing that is missing from Carr's analysis (and passed over by Whitehead) is any discussion of the positive role that AI assistance could play in addressing other cognitive deficits that are induced by resource scarcity. The work of Sendhil Mullainathan and Eldar Shafir (2014 and 2012) is instructive in this regard. It suggests that those who suffer from scarcity (income scarcity, food scarcity, time scarcity etc.) also suffer knock-on cognitive deficits (Shah et al., 2012). If a resource is scarce to you, you tend to focus all your cognitive energies on it. This has deleterious effects on your other cognitive abilities. For example, in one study, people suffering from income scarcity were found to have associated deficits in fluid intelligence (up to 15 IQ points) and executive control (impulse control, willpower etc.). This suggests that people suffering from scarcity face increased cognitive burdens, which they often find difficult to discharge. One great promise of AI assistance is that it could help to shoulder some of that cognitive burden—creating what Mullainathan and Shafir call 'slack'—which could in turn put them in a better position to address their scarcity-related problems. In other words, cognitive outsourcing through AI could redress scarcity-induced cognitive imbalances within one's larger cognitive ecology. This serves as a counterbalance to Carr's concerns about degeneration.[4]

Of course, none of this quite addresses the impact of cognitive degeneration on the intrinsic benefits of thinking deeply. There are certain cognitive tasks that I want to continue to perform (reading philosophical articles and critically reflecting on their contents being chief among them) and there are some legitimate concerns about the impact of the new cognitive ecology on my ability to do so. Here, the response to the degeneration argument should be one of cautious optimism. We need to be discerning in our approach to AI assistance. We shouldn't embrace the latest form of AI assistance

[4] I am indebted to Miles Brundage for suggesting this line of argument to me. We write about it in more detail on my webpage: https://philosophicaldisquisitions.blogspot.com/2017/05/cognitive-scarcity-and-artificial.html

without first reflecting on the nature of the cognitive task with which it assists, the likely degenerating effects of that assistance, and the possible instrumental benefits. Fortunately, there is a general principle we can apply: If the task itself is something that is intrinsically valuable, if the associated degenerating effects are likely to be widespread, and if the instrumental benefits are slight, it is probably best to avoid AI assistance. Beyond such a case, the negative consequences of degeneration can be overstated.

Determining whether the use of an AI assistant will have a damaging widespread degenerating effect will need to be assessed by reference to the commonality of the outsourced task within the individual's life, and the possible need for cognitive resiliency with respect to that task. For example, I know that budgeting and expenditure tracking is a relatively minimal part of my day-to-day life. It is not something for which I am paid or rewarded, or for which others rely on my guidance. Consequently, outsourcing the task is unlikely to have a damaging degenerating effect. Determining whether or not an activity is sufficiently intrinsically valuable to warrant avoiding AI assistance is something to be decided partly by the individual—What is most important to them? What do they find rewarding?—and partly by reference to general societal norms. I return to this latter point again in Section 5 when I discuss one set of activities that might be high on the intrinsic value scale because of societal norms, and may raise particular concerns when it comes to the use of AI assistants.

4 The Autonomy Argument

The main flaw of the degeneration argument is its over-ambitious nature. It makes the assumption that degeneration in and of itself is a bad thing when that is not always true: you need to consider the impact of that degeneration in light of the broader cognitive ecology in which a person operates. A slightly more plausible, but still reasonably broad, variant on the argument focuses on the ill-effects of AI assistance on our capacity for autonomy and responsibility.

Autonomy and responsibility are cherished values in modern liberal societies. Some of their value lies in the political and social realm: in the obligations the state (and other public actors) owe to us and that we owe to others. And some of their value lies in the personal realm. Personal goal setting and goal achievement depend on autonomy and responsibility. It is commonly believed that my happiness and self-fulfilment are best served when I pursue goals that are of my own choosing; and it is also commonly believed that the achievement and meaning I derive from my goals is dependent on my being responsible for what I do (Luper 2014; Smuts 2013). If AI assistance threatened autonomy and responsibility, it could have an important knock-on effect on our personal happiness and fulfillment.

Here is the argument someone might make:

(6) Personal autonomy and responsibility are essential for meaning, happiness and self-fulfilment.
(7) Widespread use of AI assistance undermines autonomy and responsibility.
(8) Therefore, widespread use of AI assistance undermines meaning, happiness and self-fulfilment.

Let us grant premise (6) for now. Attention then turns to premise (7): Is there any reason to believe that AI assistance poses such a threat? Yes, there is. The threat comes from two directions. First, AI assistance threatens to sever the link between what we choose and desire to do and what happens in the world around us. This can undermine personal responsibility and hence achievement. And second, AI assistance threatens to manipulate, filter or otherwise structure our choices, meaning that we act for reasons or beliefs that are not necessarily our own.

The first threat is easy to understand. Whenever AI assistance joins up with other forms of automation (e.g. the automation of physical labour), there is obviously an impact on the link between what an individual chooses or desires to do and what gets done. Suppose my partner comes home after a long day at work and asks me if I have done the vacuuming. I say 'yes, of course, my dear', beaming with pride at my industrious discharge of my household duties. It turns out that I didn't really do the vacuuming. I purchased a Roomba robot that does the vacuuming for me on an automatic cycle. So when my partner asks whether I did the vacuuming, I'm misleading her when I claim the effort as my own. I am not really entitled to feel the sense of pride and achievement that I suggested I felt. At best, I can claim responsibility for the decision to purchase and use the robot, but this puts considerable distance between me and what happens in our house on a daily basis. This might be problematic, but we need to bear in mind Whitehead's point about saving one's resources for the activities that really matter. Severing the link in certain cases seems perfectly acceptable.

The second threat is more challenging. Although the link between AI assistance and other forms of automation is real, at the moment the primary role of AI assistance is not in replacing human agency but, rather, in replacing human cognitive effort. The AIs make sense of information for us: they perform basic computations on our behalf, and then issue recommendations and suggestions to us. We are still the ones that ultimately *choose* to act on those recommendations and suggestions, but we choose from the menu of options provided by the AI. Several authors have expressed concern about this dynamic, suggesting that it is tantamount to external manipulation or coercion, and that it will ultimately corrode our autonomy.

Krakauer (2016) expresses the worry by distinguishing between two major types of cognitive artifact: the *complementary* and the *competitive*. He uses the abacus as an example of a complementary cognitive artifact. The abacus helps us to perform mathematical operations and but it does so by complementing our innate cognitive capacities. Studies of skilled abaci users show, according to Krakauer, that if you take away the abacus, their mathematical performance is not impaired. They replace the external artifact with an internal mental model. The artifact is effectively like a set of training wheels: it provides a short-term scaffold for learning. Contrast that with a competitive cognitive artifact like a digital calculator. This doesn't complement our innate cognitive capacities, it replaces them. If you take away the calculator, the human agent is no better at performing the task. AI assistants are, according to Krakauer, more like calculators than abaci and hence are likely to have this capacity-reducing effect as their usage becomes more widespread. To this extent, Krakauer is an orthodox proponent of the degeneration thesis. His explicit concern, however, is that the competition between humans and AIs will corrode

freedom and autonomy. He uses the Homeric tale of Odysseus and the Lotos Eaters to make his point:

> In Homer's The Odyssey, Odysseus's ship finds shelter from a storm on the land of the lotus eaters. Some crew members go ashore and eat the honey-sweet lotus, 'which was so delicious that those [who ate it] left off caring about home, and did not even want to go back and say what happened to them'. Although the crewmen wept bitterly, Odysseus reports, 'I forced them back to the ships…Then I told the rest to go on board at once, lest any of them should taste of the lotus and leave off wanting to get home'. In our own times, it is the seductive taste of the algorithmic recommender system that saps our ability to explore options and exercise judgment. If we don't exercise the wise counsel of Odysseus, our future won't be the dystopia of Terminator but the pathetic death of the Lotus Eaters.
>
> (Krakauer 2016)

In short, Krakauer worries that AI recommenders will make us lazy with respect to autonomy: their choices will replace our choices.

The technology critic Evgeny Morozov (2013) expresses a very similar set of concerns. He writes in particular about data-mining and predictive analytics and the impact it will have on our ability to choose for ourselves. Like Krakauer, he uses an evocative metaphor to explain his fear: the metaphor of invisible-barbed wire:

> The invisible barbed wire of big data limits our lives to a space that might look quiet and enticing enough but is not of our own choosing and that we cannot rebuild or expand. The worst part is that we do not see it as such. Because we believe that we are free to go anywhere, the barbed wire remains invisible…
>
> Thanks to smartphones or Google Glass, we can now be pinged whenever we are about to do something stupid, unhealthy or unsound. We wouldn't necessarily need to know why the action would be wrong: the system's algorithms do the moral calculus on their own. Citizens take on the role of information machines that feed the techno-bureaucratic complex with our data. And why wouldn't we, if we are promised slimmer waistlines, cleaner air, or longer (and safer) lives in return?
>
> (Morozov 2013)

So here, the concern is that the AI assistant will imprison us within a certain zone of agency. It will do all the hard cognitive work needed to come up with suggestions about what to do: we will be reduced to mere implementers of these suggestions, completely shut off from the rationale and reasons that underlie the AI's recommendations. That certainly sounds like a threat to autonomy.

But to establish whether or not it is a threat, we need to consider what the conditions of autonomy actually are. This is, unsurprisingly, a contentious philosophical question.

Raz's model of autonomy (Raz 1986) is reasonably broad, and highlights three general conditions that must satisfied if one is to exercise autonomous choice: rationality (i.e. the choice must be premised on some rationale that is comprehensible to you); optionality (you must be able to choose from an adequate range of valuable options); and independence (i.e. the choice must be free from coercion and manipulation). Other theories of autonomy point to consistency between lower-order and higher-order preferences in addition to independence from manipulation (Frankfurt 1971; Dworkin 1988). Are any of these conditions threatened or undermined by the use of AI assistance? Let's take them one by one.

The rationality condition might be threatened, depending on how it is interpreted. One consistent complaint about AI systems, particularly when they rely on machine learning algorithms, is that they are black box systems. They produce certain outputs—recommendations and suggestions—in an opaque manner: we don't know, or are unable to reverse engineer, the precise logic they use to come up with these recommendations (Burrell 2016; Danaher 2016a). This might threaten comprehensibility but how serious a threat this is depends on how extensive and deep the comprehensibility needs to be for autonomy. If an AI assistant issues one recommendation and you unfailingly and unquestioningly implement it without understanding it, then maybe the rationality of your choice is compromised. But if it gives many recommendations, you still exercise some capacity for critical reflection on the reasons for your choices. Furthermore, even in the case where it issues one recommendation, you still have some control over the decision to defer to it in the first place. The rationale/reasons underlying that deference will be comprehensible to you. It is only if the deference becomes completely automatised within your own mind that a serious problem emerges. But this then raises the paradox of internal automaticity, which I discuss in the penultimate section.

What about the optionality condition? Here, AI assistance might be able to help, not hinder autonomy. It is widely known and widely appreciated that we have far too many options available to us now. Whether it is choosing books to buy, movies to watch or songs to stream; there is too much 'stuff' out there for any one human being to process and appreciate. We need some assistance when it comes to filtering and limiting our options. Schwartz's work on the paradox of choice highlights the problem (Schwartz, 2004). Having an adequate range of options increases well-being up to a certain point, but once the range of options becomes too extensive people get 'frozen' in decision problems, unable to compute and process all the possible outcomes. This can increase feelings of regret and guilt when it comes to decision-making (Nagel 2010). Empirical evidence on the paradox of choice is somewhat mixed (Scheibehenne et al. 2010) with some studies suggesting that individuals can avoid getting 'stuck' in decision problems by adopting heuristics and other rules for filtering information. Reliance on such heuristics does, however, increase the cognitive burden when it comes to individual choices. One major advantage of an AI assistant is that it can reduce this cognitive burden by doing this filtration for you. This suggests that AI assistance can help to ensure that your choices bring about that all-important harmony between lower and higher-order preferences. By filtering options according to what it learns about your preferences, the assistant can ensure that the available options are consistent with your preferences, while at the same time reducing the opportunity for getting stuck in choices or feeling increased regret and guilt.

What about independence and manipulation? This is probably the most challenging condition. AI recommendation systems will impact on our ability to process and sort through options on our own, and this may, in turn, make us more susceptible to manipulation. As we get comfortable with outsourcing the cognitive burden of choice to the AI, we may become more trustworthy and this trust can be abused. I would argue that it is unlikely that AIs will become overtly coercive—coercion requires some explicit or implied threat that if you don't follow a recommended option you will be made worse off as a result (formally, coercion requires that you be threatened with being made worse off relative to some pre-existing baseline—see Wertheimer 1987). It is possible that AI systems will incorporate something akin to this dynamic—e.g. a health AI might threaten you with higher insurance premiums if you do not follow its recommendations—and that would clearly involve a breach of autonomy—but presumably those cases will be reasonably obvious when they arise. What is more interesting and potentially insidious—and what seems to really worry Krakauer and Morozov—is that the AI would gradually 'nudge' you into a set of preferences and beliefs about the world that are not of your own making or, as one reviewer to this paper put it, 'guilt trip' you into doing something that you would rather not do (think again of the health AI telling you that you are 2000 steps short of your 10,000 a day target). The autonomy-undermining properties of these more subtle manipulations of choice are now widely debated. Nudging, in particular, has been subject to a lot of scrutiny, with disagreement lingering as to whether it undermines autonomy or not. Nudging occurs when a choice architect designs a decision-making environment in such as way as to bias or encourage people to select certain options, e.g. putting healthy foods at eye level in a canteen in order to get people to select them over unhealthy options (Thaler and Sunstein 2009). The original claim made by Thaler and Sunstein in defence of nudges was that they could be used to satisfy desirable policy goals and while protecting autonomy. They could do so because people would always still have choices within the relevant choice architecture, even if some choices were made less attractive or more difficult. But, as Sunstein acknowledges, there is clearly a dark side to nudging and the technique could be used for nefarious purposes (Sunstein 2016). Furthermore, as the regulatory theorist Karen Yeung points out, there might be something different about the kinds of nudging that are made possible through AI assistants: they can constantly and dynamically update an individual's choice architecture to make it as personally appealing as possible, learning from past behaviour and preferences, and so make it much more likely that they will select the choice architect's preferred option. As she puts, the technology enables a kind of 'hypernudging' or nudging on steroids (Yeung 2017).

Is this problematic? Here, I think there are indeed reasons to be concerned about the impact of AI assistants, but those reasons have to be kept in perspective. In the first instance, it is important to acknowledge that our preferences and beliefs about the world are never of our own making. The projects we deem important and the options made available to us are always products of cultural and environmental forces that are beyond our control. This is, perhaps, one of the key insights of Sunstein and Thaler in their original defence of nudges: there is no perfectly neutral design for a choice architecture; every design embodies some bias or preference, even if it is not stated or appreciated. I would simply expand this observation and point out that the total set of choice architectures within any society at any given time is going to reflect a range of cultural

and historical forces that are beyond the control of the individual and that are not neutral with respect to particular options. The widespread use of AI assistants changes nothing about these basic facts. It is, consequently, short-sighted to assume that simply because AI assistants are newer and hence more salient features of our cognitive ecology that they pose a more significant threat to our autonomy. Indeed, as noted above, they may actually improve things relative to the current status quo, at least when it comes to making options more manageable.

To make the case that there is some special and significant threat you need to provide some reasons for thinking that AI assistants are radically different from what has gone before. To do this, you will need to highlight something especially problematic about the *origin/purpose* of AI-mediated influence or the *modality* of that influence. There might be reason for concern on both fronts. When it comes to the source/origin of AI-mediated influence, it is important to bear in mind that most AI assistants are constructed by commercial enterprises with their own profit-seeking agendas. We might legitimately worry that those corporations will impose preferences and options on their AI-mediated choice architectures that suit those agendas but not our own well-being. Furthermore, as a reviewer to this paper pointed out, we may worry that because there are relatively few companies controlling this technology—Google, Amazon, Facebook and Apple being the obvious culprits—the widespread use of AI assistance will lead to greater centralisation of control over choice architectures than was historically possible.

It is worth taking these concerns seriously, but again perspective is required. I think there are three reasons to be less concerned about the source/origin of AI-mediated influence than critics commonly suppose. The first is simply that certain cultural institutions have always had outsized influence over our choice architectures—e.g. religions and the State—and its not clear that the degree of power or its centralization is worse now than it was in the past. If anything the reverse might be true. The church had an outsized influence over the choice architectures of ordinary citizens in Europe for centuries and this influence expressed itself in ways that were pervasive and highly restrictive (Wu 2017). The choice architectures made possible by modern technology are more diverse and varied, and their tendency towards personalization and customisation, though costly in other ways, may often save them from the charge of being overly manipulative: they work to serve the previously expressed preferences/interests of the user and not necessarily those of the corporation controlling the technology.[5] Furthermore, centralization is a double-edged sword. Although it may increase the power of certain organisations, it also makes it easier when it comes to holding people to account when things go wrong. If AI-mediated choice architectures are hijacked to serve nefarious and manipulative purposes, then the fact that they are controlled by relatively few, well-known corporations, makes it easier to locate potentially responsible agents and hold their feet to the fire for what they have done. The recent Cambridge Analytica/Facebook

[5] I am indebted to an anonymous reviewer for suggesting the distinction between personalization and manipulation. As they pointed out, personalization also has costs, e.g. a filter bubble that serves to reinforce prejudices, that may not be desirable in a pluralistic, democratic society, but it's not clear that those problems are best understood in terms of a threat to autonomy. Cass Sunstein's #Republic (2017) explores the political fallout of filter bubbles in more detail.

scandal over the 'manipulation' of voters in the lead-up to the 2016 US Presidential Election illustrates this point. Although it clearly highlights the problems with centralised control, it also provides some reassurance that when the problem is exposed, it is possible to subject the organisation involved to critical and ultimately legal scrutiny. Finally, the centralisation of power is not an intrinsic feature of this technology. There are many smaller-scale, narrowly circumscribed, AI assistants available for use, some of which are discussed in the next section. It is possible for the individual to select a less centralised form of assistance for discrete tasks.

What about the modality of interference? Is there some distinctive threat from AI assistants on that front? The closest thing to a distinctive threat would be Yeung's concerns about hypernudging: the dynamic, constantly updated choice architecture. This is certainly worth taking seriously, but again it's not clear that this is a distinctive or significant threat. Hypernudging differs from ordinary nudging by degree and not by kind. Furthermore, the differences in degree seem to go in the direction of personalization as opposed to outright manipulation. Nevertheless, there is something to be worried about here, particularly if the hypernudging capacity is hijacked by some nefarious third party. It would seem to be impossible for an individual user to adequately address this threat by themselves—some regulatory reform/legal intervention would likely be required to deal with all the potential problems. Consideration of such reforms lies beyond the scope of this article, which is concerned with the personal ethics of AI use. Nevertheless, there is some guidance to be given at the personal level. I would argue that we should avoid a narrative of helplessness in the face of AI assistance. In the world as it is currently constituted, we are not slaves to AI assistance; we do have some residual control over the extent to which we make use of this technology. We have no legal or moral compulsion to use it, and we have our own self-judgment about the effect of certain choices on our happiness and fulfillment. This self-judgment is an authoritative source of guidance when it comes to that happiness and fulfillment: no third party can second-guess whether selecting a certain option has made us happy or not (Hare and Vincent 2016).[6] We should, therefore, not be afraid to rely on this self-judgment to either veto the influence of AI over our lives or to avoid making use of it in the first place.

So what emerges from this? Does AI assistance threaten our autonomy or not? There is no hard and fast answer. When we examine the conditions for autonomy in more detail we see that AI assistants don't necessarily pose an unusual threat to autonomy: it's unlikely that they would undermine the capacity for rational choice, and they may actually help, by pre-filtering options, to satisfy the optionality condition for autonomy. There are, however, risks of manipulation, particularly in the more subtle form of nudging or hypernudging (as opposed to the overt form of coercion). Some regulatory intervention and reform, such as the reforms under the new GDPR, will be needed to adequately address all of those risks. Nevertheless, we are not individually powerless in the face of such threats: people need to be made more cognizant of this. One of the purposes of this article is to do exactly this and to formulate principles that individuals could use to address the risks. A

[6] As Hare and Vincent point out, while humans may be bad at predicting whether a future option will makes us happy, our judgment as to whether a chosen option has made us happy is, effectively, incorrigible. Nobody knows better than ourselves. It is to this latter type of judgment that I appeal in this argument.

suggested principle here would be to get the person think about the decision-making context in which the AI assistant is being used, and to consider the source/motivations underlying the creators of the assistant: If you are using AI in a choice-context where there is not an overwhelming number of options, and if the commercial motivations, or centralised power of the organisation behind the app are abhorrent to you, then you probably should avoid using it. If the opposite is the case, there is likely nothing to fear and much to gain. In the intermediate case, where there are overwhelming options but also abhorrent motivations/centralising power, more judgment will be required as the individual will need to trade-off the risks involved.

5 The Interpersonal Communication Argument

This brings us to the final objection (or set of objections). The preceding objections tended toward the general: the degeneration of cognitive faculties and the undermining of autonomy and responsibility. The critical evaluation in each instance tended toward the specific: generalised objections are too vague to work; we need to pay attention to the specific ecological context in which AI gets used and the impact it has on cognitive ability, freedom and responsibility in those contexts. Now, I want to look at an objection that is tailored to a more narrowly conceived set of contexts, while still retaining some degree of generality. I do so partly because I think the objection is interesting in its own light, and partly in order to answer a challenge that was left lingering from the initial discussion of Carr's degeneration argument. If you recall, I noted there that in order to convince us to take degeneration seriously, Carr either needed to show that it had some pervasive negative effect on our cognitive ecology (which I argued he could not do), or that it would knock-out or eliminate some set of abilities with really high intrinsic value, and whose loss would make our lives much worse. Is there any such risk?

Evan Selinger's work is instructive in this regard. In a series of articles and op-eds (Selinger 2014a, 2014b, 2014c; and Selinger and Frischmann 2016), he has articulated ethical objections to the use of AI that apply, in particular, to interpersonal communications. He focuses specifically on the negative impact of automated communication services. The essence of all these services is that they allow you to send text messages to your friends and intimate partners on preordained or random schedules, without the need for you to draft the messages at the particular moment that they are sent. Sometimes this is because you will have written the messages long in advance; sometimes it will be because you will have selected messages from a menu of options provided by the automated service; and sometimes it will be because the service drafts, selects and sends the message for you.

There are several services that do this. Google's *Allo* is a smart messaging platform that uses Google's general *Assistant* AI software to learn about your messaging style, suggest messages to you, and automate sending and responding to messages. More specific services include apps like *Romantimatic* and *Bro App*, which are designed to send affectionate messages to romantic partners. *Romantimatic* seems well-intentioned, and is primarily designed to remind you to send messages (though it does include pre-set messages that you can select with minimal effort). *Bro App* is probably less well-

intentioned (though it may be intended as a joke). It is targeted at men and is supposed to allow them to spend more time with their 'bros' by facilitating automated messaging. While these examples are interesting, it is important not to get too bogged down in their details. Apps of this sort pass in and out of existence frequently. What is popular today may no longer be popular tomorrow. The more important thing is the general phenomenon of automated messaging, which is made possible by advances in AI. Should we worry about this phenomenon? Selinger thinks we should, at least in some interpersonal contexts. While he fleshes out his concerns in several ways in his writings, two specific concerns stand out.

The first is a worry about deceptiveness/inauthenticity:

> …the reason technologies like BroApp are problematic is that they're deceptive. They take situations where people make commitments to be honest and sincere, but treat those underlying moral values as irrelevant — or, worse, as obstacles to be overcome.
>
> (Selinger 2014a)

When sending messages to romantic or intimate partners, you should be sincere in what you say. If you say, 'I'm thinking about you' or 'I miss you' or 'I love you', you should mean what you say. If you haven't actually drafted the message, or you have automated its communication in order to spend more time with your friends, then you are not being sincere. To put this into argumentative form:

(9) It is a bad thing to be deceptive in your interpersonal relationships.
(10) Automated messaging services encourage deceptive interpersonal communications.
(11) Therefore, these apps are bad things.

This is a relatively weak objection to the use of automated messaging. Premise (9) would seem to be flawed. There may be contexts in which a degree of deceptiveness is desirable in interpersonal relationships. The so-called 'white lies' that we tell to keep our partners happy may often be justifiable and may help to sustain rather than undermine a relationship. And premise (10) is definitely problematic insofar as the mere automation of a message does not make the sentiment or meaning deceptive. Deceptiveness is usually taken to denote an active intent to mislead another as to the context or the truth of what you are saying. It's not clear that automated messaging always involves that kind of active intent. Something like Bro App may do so, but this is a rather exceptional service. When you focus on the more general phenomenon, there is no reason why someone could not set up a pre-scheduled list of messages that sincerely and truthfully conveyed their feelings toward another person. Suppose at the start of the week I ask Siri or Google Assistant to text my partner at three random intervals with affectionate messages. Suppose the messages are sent and received at those times, and my partners feels good about them. Am I being deceptive or inauthentic simply because I dictated it all in advance? Not necessarily. As long as I retain a principal-agent type relationship with the automated messaging service, there is no

reason to think that the actions performed by the service are not accurately communicating my true feelings towards my partner.[7]

This brings us to Selinger's second concern about this type of technology. Beyond deceptiveness, he suggests that there are some interpersonal communications in which the value of the communication lies in the fact that it is an immediate, deliberate and conscious representation of how you feel about another person. In other words, that sometimes the real value of receiving an affectionate message from a loved one, lies not in the content of the message or the effect it produces, but rather in the fact that the other person is really thinking about you at that moment—that they are being intentional and 'present' in the relevant communicative context. The problem is that the very logic of these apps—the automated outsourcing of communications—serves to corrode this 'real' value. The apps are telling us that it is no longer important to be conscious and present in our communications; they tell us that what matters is content and effect; and that you can produce the required content and effect without being immediately present:

(12) The real value of certain types of interpersonal communication is that they are immediate, conscious and intentional representations of how we feel.
(13) Automated messaging services (by their very nature) create communications that are not immediate, conscious and intentional.
(14) Therefore, automated messaging services undermine the real value of certain types of interpersonal communication.

This is a more formidable objection because it gets to the heart of what AI assistants do (namely: redistribute cognitive labour). It is as close to an in-principle objection to the use of AI assistants, in a certain context, as you are likely to get. But there are several limitations to bear in mind.

First, this objection cannot be easily generalised. It certainly applies beyond the communicative context, but even then it only applies to those cases in which the primary value of an act or performance lies in the fact that it is performed with immediate, conscious intention. This brings us back to the trade-off between intrinsic vs. instrumental values within our cognitive ecology that was raised earlier when discussing the degeneration problem. If a cognitive task derives most of its value from its extrinsic (instrumental) properties, then the objection does not apply. For example, the value of dividing up a bill at a restaurant does not lie in the conscious performance of the arithmetical operation; it lies in getting the right result. Outsourcing this activity seems unobjectionable. The same goes for many interpersonal communications. If my partner asks me to text her when I arrive at my destination after a long journey, the value of the message lies in conveying truthful information about my safe arrival, not in my consciously sending it at that particular moment in time. So whenever we have a case in which the value of the performance is instrumental, we can use the AI assistant without great concern.

The tricky question is what we should do when the value of a performance is partly instrumental and partly intrinsic and there is no clearly asymmetric distribution of value across the instrumental and intrinsic components. The value of saying 'I love you'

[7] As a reviewer points out, it may be impossible for interpersonal communication to ever adequately capture one's true feelings. This may well be right but if so it would seem to be a problem for both automated and non-automated communications alike.

might sometimes lie in the fact that it is an immediate conscious expression of your feelings (e.g. if you are in a face-to-face setting and your partner says 'I love you', sending a text message hours later that reciprocates might not cut it). At other times, the value might lie more in its sincerity and the effect it produces in your partner, not whether it tracks your immediate conscious affect. On still other occasions, the value might oscillate back and forth between these two poles.

In those latter cases, the solution to Selinger's ethical objection would be to ensure that the AI assistant does not completely dominate the performance of the activity. In other words, to ensure that we allow ourselves to take control and be involved in the immediate conscious performance of the activity on at least some occasions (or on those occasions when it seems most appropriate). This is where Carr's concerns about degeneration and cognitive resilience have most force. It is important to maintain some cognitive resilience so that we can be occasionally involved in those activities where immediate conscious intention is sometime desirable. This, however, leads to an important paradox.

6 The Paradox of Internal Automaticity

Selinger's criticisms of AI are interesting because they highlight the value of immediate conscious intention in some activities. But in doing this, they also force us to confront the paradox of internal automaticity. Throughout this article, we have been assuming that automation involves outsourcing cognitive performance to a machine (AI) that is external to your body and mind. But this assumption is not quite right. Automation can happen inside the body too. When you first learn how to do something (like ride a bike or drive a car or talk in front of an audience), it requires intense, occurrent, conscious effort: you have to think and represent each stage of the performance in order to get it right. But once the performance is mastered, it becomes automatic: a subconscious routine that you can easily perform without the intense conscious effort. This is referred to as *automaticity* in the psychological literature. Some of the objectors to AI outsourcing are familiar with this phenomenon. Nicholas Carr, for one, pairs his discussion of cognitive degeneration with a sensitive account of internal automaticity. But he has no qualms or concerns about internal automaticity.

This is, prima facie, paradoxical. Both internal automaticity and external automation involve the bypassing of immediate conscious intention. If the latter is problematic, then it seems like the former should be too. If, when my partner says 'I love you', I automatically and without thinking say 'I love you' back, this would seem to be just as problematic as if I got an app to send the message on my behalf (if we assume, *arguendo*, that the value of the communication lies primarily in its immediate conscious performance rather than the effect it produces). But the reality is that reflexive, unthinking, internally automated actions of this sort are commonplace. We often zone out of what we do because it is routine. Those who think that this is fine, but that external automation is not owe us some explanation of why this is the case. That's the paradox of internal automaticity.

There are at least three ways to address the paradox. The first is simply to concede defeat and accept that internal automaticity is just as much a problem as external automation, at least in those contexts where immediate conscious participation in an act is ethically preferred. This does not weaken the objection to the use of personal AI

assistants, but it does suggest that the objection lies within a more general category of objection, namely: thoughtless engagement in certain activities. Personal AI assistants are then problematic to the extent that they encourage or foster this thoughtlessness, but lots of other things are too. The second way to address the paradox is to suggest that internal automaticity is less of problem due to its *locus of control*. This gets us back into the earlier debate about autonomy and manipulation. It could be that the internally automated behaviours are more under the control of the biological agent than external automating devices (which may be controlled by corporations with ulterior motives). This allows the objector to maintain some principled distinction between the two phenomena, but it's not clear how persuasive the distinction is. Many internally automatised behaviours (e.g. default greetings and utterances) are the product of external cultural forces acting on the biological agent. These forces may also have ulterior motives or unwelcome consequences, e.g. automatised deference to one's perceived social superiors is common in hierarchical societies and may serve to perpetuate the inequalities in such societies. A third way to address the paradox is to appeal to a conservative bias, i.e. to say that because internal automaticity has deeper biological and cultural roots, it is less morally problematic than external automation. Indeed, depending on how we interpret the conservative bias, we may even be inclined to presume that internal automaticity has some important intrinsic or instrumental value. As one reviewer to this article put it, internal automaticity is an essential conserved evolutionary resource: it is a way of adapting and learning behaviour at a relatively low cost. We have reason to be grateful for it. This conservative bias is fine insofar as it goes, but it is unlikely to appeal to those who demand some rationale or explicit justification in their applied ethical reasoning.

The point here is not to suggest that the paradox of internal automaticity is fatal to objections to personal AI assistance. The point, rather, is to suggest that once we are aware of it, some of the objections to AI assistance are a little bit blunter, since we already live with the reality they seem to fear.

7 Conclusion

This article has examined the ethics of personal AI assistants. It has argued that we should look at the use of such technology through an ecological lens: as something that changes the cognitive ecology in which we operate and that requires careful consideration of the effects within that ecology, particularly in terms of the trade-off between instrumental and intrinsic value. It has assessed some paradigmatic ethical objections to the personal use of AI assistants, arguing that while none of them amounts to knock-down objection to the practice, they do raise important concerns. Responding to those concerns allows us to draft an initial framework for thinking about the sensible, ethical use of personal AI assistance. The framework comes as a series of ethical risk/reward guidance principles that ask users to pay attention to potential risks in certain contexts and to tailor their use of AI assistance in response to those risks:

Degeneration Risk/Reward When using AI assistants, one should be aware of the potential for cognitive degeneration with respect to an assisted task. If the task is one whose primary value is instrumental, and if the risk of being decoupled from the AI

assistant is minimal, then this may not be a problem. But if the primary value of the task is intrinsic, and/or the risk of decoupling is high, it is probably best to cultivate one's own cognitive capacities for the task.

Autonomy Risk/Reward When using AI assistants, one should be aware of the potential threat to autonomy due to (a) the fact that AI can sever the link between choice and result and (b) the fact that AI assistants can design the choice architecture in which we make decisions. But one should bear in mind that AI assistants do not pose a unique and special threat to autonomy—many cultural and environmental forces already structure choice architectures in problematic ways—and that there is a danger in adopting a narrative of helplessness in relation to such threats. One should, consequently, approach personal AI assistants with the same discernment with which we approach existing threats. Furthermore, one should bear in mind that if the assistance comes in a domain with a lot of options, then the pre-selection may be advantageous and autonomy-enhancing.

Interpersonal Virtue Risk/Reward When using AI assistants to mediate interpersonal relationships, one should be aware that the primary value of some interpersonal actions comes from immediate, conscious engagement in the performance of that action. To the extent that AI assistants replace that immediate, conscious engagement, they should be avoided. Nevertheless, in many other cases, the value of interpersonal actions lies in their content and effect; in these cases, the use of AI assistants may be beneficial, provided they are not used in a deceptive/misleading way.

This is, of course, very much a first draft. The intention would be for these risk/reward principles to be borne in mind by users of the technology as they try to make judicious use of them in their lives. But the principles could also be of use to designers. If they wish to avoid negatively impacting on their user's lives, then considering the effect of their technologies on cognitive capacity, autonomy and interpersonal virtue would be important. Further elaboration of this framework is certainly required, and more guidance on which types of activity derive their value from immediate conscious engagement or the situations/abilities that would be most affected by decoupling, and hence in need of some resiliency, would be desirable, but it is hoped that this provides a platform on which others can build.

References

Burgos, D, Van Nimwegen, C, Van Oostendorp, H. and Koper, R. (2007). Game-based learning and immediate feedback. The case study of the Planning Educational Task. *International Journal of Advanced Technology in Learning* Available at http://hdl.handle.net/1820/945 (accessed 29/11/2016).

Burrell, J. (2016). How the machine thinks: Understanding opacity in machine learning systems. *Big Data and Society*. https://doi.org/10.1177/2053951715622512.

Carr, N. (2014). *The glass cage: Where automation is taking us*. London: The Bodley Head.

Crawford, M. (2015). *The world beyond your head*. New York: Farrar, Strauss and Giroux.

Danaher, J. (2016a). The threat of algocracy: Reality, resistance and accommodation. *Philosophy and Technology, 29*(3), 245–268.

Danaher, J. (2016b). Why internal moral enhancement might be politically better than external moral enhancement. Neuroethics. https://doi.org/10.1007/s12152-016-9273-8

Dworkin, G. (1988). *The theory and practice of autonomy*. Cambridge: CUP.

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

Frischmann, B. (2014). Human-focused Turing tests: A framework for judging nudging and the techno-social engineering of humans. *Cardozo Legal Studies Research Paper No. 441* - available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2499760 (accessed 29/11/2016).

Giublini, A., & Savulescu, J. (2018). The Artificial Moral Advisor. The 'Ideal Observer' meets Artificial Intelligence. *Philosophy and Technology, 31*(2):169–188.

Hare, S., & Vincent, N. (2016). Happiness, cerebroscopes and incorrigibility: Prospects for Neuroeudaimonia. *Neuroethics, 9*(1), 69–84.

Heersmink, R. (2015). Extended mind and cognitive enhancement: Moral aspects of extended cognition. *Phenomenal Cognitive Science*. https://doi.org/10.1007/s11097-015-9448-5.

Heersmink, R. (2013). A taxonomy of cognitive artifacts: Function, information and categories. *Review of Philosophical Psychology, 4*(3), 465–481.

Kelly, S., & Dreyfus, H. (2011). *All things shining*. New York: Free Press.

Kirsh, D. (2010). Thinking with external representations. *AI and Society, 25*, 441–454.

Kirsh, D. (1995). The intelligent use of space. *Artificial Intelligence, 73*, 31–68.

Krakauer, D. (2016). Will AI harm us? Better to ask how we'll reckon with our hybrid nature. *Nautilus* 6 September 2016 - available at http://nautil.us/blog/will-ai-harm-us-better-to-ask-how-well-reckon-with-our-hybrid-nature (accessed 29/11/2016).

Luper, S. (2014). Life's meaning. In Luper (Ed.), *The Cambridge Companion to Lie and Death*. Cambridge: Cambridge University Press.

Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data and Society*. https://doi.org/10.1177/2053951716679679.

Morozov, E. (2013). The real privacy problem. MIT Technology Review. Available at http://www.technologyreview.com/featuredstory/520426/the-real-privacy-problem/ (accessed 29/11/16).

Mullainathan, S. and Shafir, E. (2014) Freeing up intelligence. *Scientific American Mind* Jan/Feb: 58–63.

Mullainathan, S., & Shafir, E. (2012). *Scarcity: The true cost of not having enough*. London: Penguin.

Nagel, S. (2010). Too much of a good thing? Enhancement and the burden of self-determination. *Neuroethics, 3*, 109–119.

Nass, C. and Flatow, I. (2013) The myth of multitasking. *NPR: Talk of the Nation* 10 May 2013 - available at http://www.npr.org/2013/05/10/182861382/the-myth-of-multitasking (accessed 29/11/2016).

van Nimwegen, C., Burgos, D., Oostendorp, H and Schijf, H. (2006). The paradox of the assisted user: Guidance can be counterproductive. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 917–926.

Newport, C. (2016). *Deep Work*. New York: Grand Central Publishing.

Norman, D. (1991). Cognitive artifacts. In J. M. Carroll (Ed.), *Designing interaction: Psychology at the human-computer interface*. Cambridge: Cambridge University Press.

Ophir, E., Nass, C., & Wagner, A. (2009). Cognitive control in media multitaskers. *PNAS, 107*(37), 15583–15587.

Pinker, S. (2010). The cognitive niche: Coevolution of intelligence, sociality, and language. *PNAS, 107*(Suppl 2), 8993–8999.

Plato. *The Phaedrus*. From *Plato in Twelve Volumes*, Vol. 9, translated by Harold N. Fowler. Cambridge, MA, Harvard University Press; London, William Heinemann Ltd. 1925. Available at http://www.english.illinois.edu/-people-/faculty/debaron/482/482readings/phaedrus.html (accessed 29/11/2016).

Raz, J. (1986). *The morality of freedom*. Oxford: OUP.

Russell, S. and Norvig, P. (2016) *Artificial intelligence: A modern approach* (Global 3rd edition). Essex: Pearson.

Sandel, M. (2012). *What money can't buy: The moral limits of markets*. London: Penguin.

Scheibehenne, B., Greifeneder, R., & Todd, P. M. (2010). Can there ever be too many options? A meta-analytic review of choice overload. *Journal of Consumer Research, 37*, 409–425.

Scherer, M. (2016). Regulating artificial intelligence systems: Challenges, competencies and strategies. *Harvard Journal of Law and Technology, 29*(2), 354–400.

Schwartz, B. (2004). *The paradox of choice: Why less is more*. New York, NY: Harper Collins.

Selinger, E. and Frischmann, B. (2016). The dangers of Smart Communication Technology. *The Arc Mag* 13 September 2016 - available at https://thearcmag.com/the-danger-of-smart-communication-technology-c5d7d9dd0f3e#.3yuhicpw8 (accessed 29/11/2016).

Selinger, E. (2014a). Today's Apps are Turning us Into Sociopaths. WIRED 26 February 2014 - available at https://www.wired.com/2014/02/outsourcing-humanity-apps/ (accessed 29/11/2016).

Selinger, E. (2014b). Don't outsource your dating Life. *CNN: Edition* 2 May 2014 - available at http://edition.cnn.com/2014/05/01/opinion/selinger-outsourcing-activities/index.html (accessed 29/11/2016).

Selinger, E. (2014c). Outsourcing Your Mind and Intelligence to Computer/Phone Apps. *Institute for* Ethics *and Emerging Technologies* 8 April 2014 - available at http://ieet.org/index.php/IEET/more/selinger20140408 (accessed 29/11/2014).

Shah, A. K., Mullainathan, S., & Shafir, E. (2012). Some consequences of having too little. *Science, 338*, 682–685.

Slamecka, N., & Graf, P. (1978). The generation effect: The delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory., 4*(6), 592–604.

Smuts, A. (2013). The good cause account of the meaning of life. *Southern Philosophy Journal, 51*(4), 536–562.

Sunstein, C. (2016). *The ethics of influence*. Cambridge, UK: Cambridge University Press.

Sunstein, C. (2017). *# Republic: Divided democracy in an age of social media*. Princeton, NJ: Princeton University Press.

Thaler, R., & Sunstein, C. (2009). *Nudge: Improving decisions about health, wealth and happiness*. London: Penguin.

Wertheimer, A. (1987). *Coercion*. Princeton, NJ: Princeton University Press.

Whitehead, A. N. (1911). *An introduction to mathematics*. London: Williams and Norgate.

Wu, T. (2017). *The Attention Merchants*. New York: Atlantica.

Yeung, K. (2017). 'Hypernudge': Big data as a mode of regulation by design. *Information, Communication and Society, 20*(1), 118–136.