This is a preprint version of a paper forthcoming in *AI & Society*. Please refer to the published version for citation purpose.

François Kammerer
Université catholique de Louvain

Self-building technologies

Abstract: On the basis of two thought experiments, I argue that *self-building technologies* are possible given our current level of technological progress. We could already use technology to make us instantiate selfhood in a more perfect, complete manner. I then examine possible extensions of this thesis, regarding more radical self-building technologies which might become available in a distant future. I also discuss objections and reservations one might have about this view.

Introduction[1]

The development of artificial intelligence and the possibility of cognitive enhancement that such a development offers – for example, through putative *merging* with AI – has recently raised concerns about the impact such enhancements could have on our *selves*. Could the process of enhancing ourselves by *merging* with AI lead to a *loss* of selfhood, or even to the *destruction* of our own selves as such – for example, because the transition from carbon-based to silicon-based cognition that such a process implies does not allow for the upholding of consciousness or personal identity? (Schneider, 2009, 2019; Schneider & Mandik, 2018) These concerns are often balanced with the potential *gains* we can expect from merging with AI, in terms of intelligence, well-being, power or lifespan. Consequently, the question that arises sometimes seems to be: "should we merge with AI, and take the risk of sacrificing or destroying our own selves, in order to make huge gains in intelligence, well-being, etc. – or not?" (Agar, 2010, 2012, 2014; Levy, 2011)

Although I take these concerns to be legitimate, I suspect their discussion sometimes ignores some potential uses of AI technology and cognitive enhancement: uses which correspond to what I call *self-building technologies*. Cognitively enhancing ourselves with the help of AI technology could not only make us gain intelligence, well-being, power or lifespan. It could also make us become *more genuine selves* – by increasing the control we have on our behavior, as well as the coherence and the transparency of our cognitive and emotional lives. My goal here is to argue that such self-building technologies are possible (and arguably likely to be created), even given our *current state* of technological progress. Moreover, future technological progress might make *radical* versions of such technologies available, which could radically change the kind of beings we are.

I describe two examples of possible technologies, which could already be implemented now (or in a near future) given the current state of technological progress (§1). Second, I argue that these technologies would count as *self-building technologies*, and I speculate on the possibility of future, more radical self-building technologies (§2). Third, I examine objections to the view that my examples are cases of genuine self-building technologies (§3). Finally, I examine and I discuss some more general reservations one could have regarding self-building technologies (§4).

1. Two examples of (imaginary) technologies

I will start by describing two examples of a *possible* use of technologies, which I will later argue are examples of self-building technologies. The description of these technologies will be made without using the concept of *self*, which is why I keep the definition of the notion of self for the next section. These technologies are such that agents could be motivated to use them because they would meet their

1 I would like to thank Peter Clutton, Keith Frankish, Julias Haas, Ben Henke, Colin Klein, Ignacio Quintana, Eric Schwitzgebel for their helpful comments, as well as the audience at the Ernst Mach Workshop VIII in Prague. I also thank Sonia Paz-Higgins for her help.

preexisting social and psychological needs, and *not specifically because these technologies would be self-building.* Importantly, the two examples I will describe do not presuppose a level of technological progress distant from ours. As far as I know, these technologies *could already be implemented* now (or at least in a near future).

A/ Pr. Truffle the implicit racist and iDiversity®.[2] Pr. Truffle is a white philosophy professor who works in a US university. In public and in private, he professes the intellectual equality of all races. He has studied the issues in detail and, when given the opportunity, he sincerely and competently argues in favor of egalitarian views of races. However, when it comes to his spontaneous reactions and implicit judgments, he is almost systematically racist. When students ask questions in class, he cannot help but think that some questions sound smarter than others – and white students almost always seem to ask smarter questions than black students (even when they don't). When he reads students' essays, he cannot help but think that some essays reveal more philosophical depth than others – and white students almost always seem to write deeper essays than black students (even when they don't). When a black student submits a particularly brilliant essay, he is more surprised than if a white student does so, and he is also more likely to suspect cheating or plagiarism. When he sits on a hiring committee, white applicants systematically seem smarter to him than black applicants (even when they aren't).

Pr. Truffle is himself rather gifted at self-observation, he is intellectually very honest and he has read extensively on implicit bias. Consequently, *he is perfectly aware* of his racist bias – and does not indulge in it. He actively tries to reform and to counter his own bias. He reads papers on the topic and attends workshops and training sessions on diversity. When he grades papers or talks with students, he sometimes tries to be extra-charitable with black students. However, this often backfires, as it leads him to act in an unnatural and patronizing manner, quite distinct from the kind of racial fairness he is really aiming at. Moreover, it is at any rate impossible for him to constantly make such an effort to counter his bias. Most of the time, his unguarded behavior and judgments simply end up merely reflecting his racist bias.

We are in Fall semester 2023, and Pr. Truffle has been asked by the head of the philosophy department to launch a new program of online philosophy courses. Pr. Truffle is allowed to work from home – which is perfect for him, as he hates commuting. During this academic year, his only interactions with students and colleagues will be *via* the internet. His teaching activity will go like this: he will make short instructional videos for the students and post them on the university website together with the reading list. Students will then send in their questions and submit their essays on the website.

When he sets his online account on the university website, he is offered the possibility to use iDiversity®, a software recently developed by a start-up hosted at the university. Pr. Truffle freely chooses to use it (the use of the app is not mandatory, and no one outside of him will know whether or not he uses it). iDiversity® is an application that works as an add-on to his online account on the university website. It does three things:

(a) It systematically anonymizes the emails, questions and essays he receives from students, as well as the cover letters and any CV he receives from applicants.
(b) It scans the emails and essays written by students to detect racially-laden content, and randomly replaces such content. For example, stylometric studies made on big data might show that certain words or expressions are more often used by black students than white students, or the other way around. iDiversity® randomely replaces some of these by expressions typically used by other groups, so as to make the ethnicity of the student as difficult to perceive as possible for the professor.

---

2 The case of Pr. Truffle (prior to his use of iDiversity®, which I describe below) is essentially similar to the case of *Juliet, the implicit racist*, described by Eric Schwitzgebel (Schwitzgebel, 2010) – which is here my main source of inspiration. Well-read readers might also have spotted Molière's *Tartuffe* as a secondary source of inspiration.

(c) It scans the videos, the emails and the comments which are about to be sent by the professor in order to detect racist (or racially-laden) content and formulations, and systematically suggests alternatives. For example, it might be that Pr. Truffle tends to choose stereotypically "white" names when he provides philosophical examples in his videos or his comments. iDiversity® can signal that to the professor, and suggest alternatives.

To sum-up: iDiversity® modifies the *input* Pr. Truffle receives from students and applicants *via* his online account on the university website, in order to make their ethnicity as difficult to perceive as possible (while changing as little as possible the content of the emails and essays). It also draws the attention of the professor to potentially racist aspects of his *output*, and gives him the opportunity to modify problematic content that has escaped his attention.

What can Pr. Truffle expect from iDiversity®? Ideally, it could help him modify his unguarded behavior and spontaneous judgments, so as to make them *less racist* – and much more efficiently than when he simply relied on his own unaided efforts. Of course, Pr. Truffle, while using iDiversity®, would arguably *retain his implicit racist dispositions and biases*. However, the modification of the *input* makes it so that the conditions of manifestation of these racist dispositions would obtain less often (as Pr. Truffle would have a harder time knowing the ethnicity of the person he interacts with), if at all – at least in a professional context. Moreover, even when these dispositions manifest, iDiversity® *partially* prevents them from having an impact on the interaction by drawing Truffle's attention to the problematic aspects of his *output* – giving him the opportunity to reflectively change it.

B/ Emma the inconstant wife and iFidelity®.[3] Emma is a young stay-at-home heterosexual wife, who is deeply in love with her husband Charles. They live together in the rich neighborhood of a big city in Europe. Charles is a very successful medical doctor in a private hospital; he is intelligent, sensitive, nice, meek and generous. He makes Emma happy, and he would even make her *perfectly* happy if it wasn't for one thing: his looks. Emma, unfortunately, happens to be very sensitive to male beauty, while Charles is universally judged to fall on the "ugly" side of the spectrum.

When they are together at home, Charles' poor looks do not particularly bother Emma. However, whenever she goes out and sees an attractive man, she cannot help feeling a strong attraction for the person to whom the face belongs. She starts fantasizing sexually and sentimentally, wishes ardently that the handsome man would notice her. She frequently has crushes on attractive men she meets, even if she meets them for a very short amount of time, and it seems that handsome men are everywhere – which means she has *many* crushes.

Emma loves her husband, and they both value monogamy and fidelity. She never seriously thinks about having an affair – let alone about divorce. However, this also implies that, whenever she sees one of these men she fancies, she feels deeply frustrated, as she knows her desires will not be satisfied. She also cannot help but deeply resent her husband for not being as attractive as these other men (even though she is angry at herself for being so irrational and unfair!), and for preventing her, by his very existence, from pursuing her fantasies. In these moments, she starts being angry at Charles for no apparent reason. She starts random fights that she later regrets. This issue, she thinks, seriously endangers her happiness, as well as her marriage.

---

3 The case of Emma is partly inspired by a story also written by Eric Schwitzgebel (Schwitzgebel, 2019), titled "My daughter's rented eyes". In this story, a young blind girl is provided with artificial eyes. The Eye & Ear Company renting the eyes gives a low rental price; in exchange, they require the parents of the girl to accept some degree of control by the company on the young girl's visual input. At the start, the modification consists in making certain visual stimuli more salient (e.g. the logos of companies who partner with Eye & Ear Company); each update comes with more modifications, which end up giving immoderate power to the company. As one might have noticed, the case of Emma is also partly inspired by Flaubert's *Madame Bovary*.

Emma knows that she does not really *need* a more attractive husband, and that her resentment and her anger only arise when she *sees* attractive men around. She remembers that, two years ago, she went on summer vacation with Charles for two weeks in a little village in Scotland, only populated by a few old couples. She never felt happier that during these two weeks: free from the excruciating desires created in her by the sight of attractive strangers, she could fully appreciate the qualities of her husband, deprived of any resentment. She often wishes they could live in a place like that – without anyone around to capture her imagination.

In May 2024, Emma turns 30. For her birthday, she receives augmented-reality glasses: discreet, elegant glasses that you can wear daily, and which directly provide visual extra-information about the environment to their users. On the app-store of her glasses, she sees that a new app just came out: iFidelity®. She decides to buy this app, and starts using it. iFidelity® does the following things:

(a) Using some facial recognition device, it *detects* which human faces visible in the environment are likely to be judged *attractive* by the user (a calibration session of a few hours for each user is first needed, in which users have to rate the attractiveness of thousands of faces).
(b) It then slightly *deforms* these human faces as they are shown to the user, in order to make them seem, through the glasses, *less attractive* than they are. The deformation is done in a very natural and plausible way: certain small facial features are simply very slightly moved, made bigger or smaller, so that the result is a very plausible, mildly unattractive human face. The change is made always in the same exact way for each individual face, so that the process creates no real issue when it comes to *recognizing* different people through time. The result is simply that all attractive people seem less attractive, or even not attractive at all (depending on which option you choose) – although you can of course decide to "opt out" some particular people from the visual deformation process (partners, friends, family, etc.).[4]

To sum-up: iFidelity® modifies the visual input delivered to Emma, to stop her having visual experiences of attractive male faces. At the same time, it modifies the rest of her visual input as little as possible, so that she can still do most of the things she currently does thanks to her ability to visually perceive faces – recognize people through different encounters, describe their appearance to others, etc.

What can Emma expect from iFidelity®? She hopes that it would help her suppress the sexual and sentimental fantasies she cannot help having about the good-looking men she sees – and, consequently, the resentment and the anger against her husband these frustrated desires tend to create – by changing the visual *input* she gets from looking at men's faces. Of course, Emma, while she uses iFidelity®, would keep her *dispositions to appreciate male beauty*, as well as her disposition to be *moved* by male beauty on a sexual and sentimental level. However, the modification of her visual input by the app would make it so that the conditions of manifestation of these dispositions would obtain much less often, as Emma would simply not *see* handsome male faces anymore, even though she will *know* that they still exist out there.[5]

---

4 We could also imagine an option in which the app does the exact opposite for some selected faces, so that these faces consistently appear *more beautiful* than they are. Emma could then chose to make her husband look better to her own eyes. Given the way in which the technology is implemented though, there will be moments where she takes her glasses off, so that the real face of her husband appears (for example, in intimacy), which would partly defeat the purpose, and create some uncanny situations. However, permanent lenses would maybe make that practicable.

5 Eric Schwitzgebel pointed out, while reading about this example, that the case of Emma has some similarity to the situation described in Ted Chiang's short story, "Liking what you see" (Chiang, 2002). In this story, Chiang imagines people who wear transcranial devices disabling the part of people's brains that make beauty judgments about people, which notably allows to end discriminations against less attractive people. Thanks to Eric Schwitzgebel for the pointer.

Numerous other similar examples could be developed[6], but I suggest focusing on these two: *iDiversity*® and *iFidelity*®. I take it that both these technologies could technically be developed now, or in a near future.[7]

2. Self-building technologies

A/ Selfhood

Before I explain why iDiversity® and iFidelity® should count as self-building technologies, I need to say a bit more about what I take *selves* to be. So, what is it to be a self?

First, I consider "self" to be more or less an equivalent of "person" (*minus* the stronger moral/normative connotations of the term "person"). Second, when it comes to the question "what is a self?" and "are there selves?": as a matter of presupposition, I rule out primitivist and anti-naturalist conceptions of selves, as well as nihilist "no-self" views, according to which selves simply do not exist. (I will briefly get back to views of this kind in the next section).

I use a naturalist and realist conception of the *self*, according to which *selves* are natural entities: something is a self when something is a psychological creature, endowed (notably) with certain psychological features and capacities. Amongst the mental features and capacities which are required in order for something to be a self, are the following:

(a) In order to be a self, a creature must have a certain degree of psychological coherence.[8] A creature who, say, holds wildly incoherent beliefs, reactions, emotions, memories, cannot be properly said to be *a* self in a unified sense – although she might be seen as the host of two, or more, selves.

(b) In order to be a self, a creature must have the ability to self-represent with a certain degree of transparency[9]. A creature who is not able to self-represent, or only self-represents in radically incorrect ways, or only self-represents in completely indirect ways (not more efficiently, directly or accurately than it represents other creatures, and without distinctively first-personal representations), cannot be properly said to be a self.

---

6 For example, think of *Jerome, the righteous-but-lazy consumer*. Jerome, for ethical and political reasons, would really like *not to buy stuff made in dictatorships*, but he is simply too lazy or too forgetful to systematically *check* the origin of all the products he buys online. Jerome could then use *iConsumer*®, an add-on to his internet browser: *iConsumer*® could automatically hide all the products made in certain countries (countries that fare badly when it comes to human rights, say) from the list he is presented with whenever he goes online shopping.

7 I am unsure, for example, whether facial-recognition technologies, and face-modification technologies, are already developed enough to make something like iFidelity® possible, although they have made tremendous progress in recent years – see for example the recent worries about the increasing difficulties to detect "deepfake" videos (Güera & Delp, 2018).

8 The insistence on psychological coherence (particularly of *diachronic psychological coherence*, or *continuity*) as a condition for selfhood (or personhood) is often associated with the Lockean conception of the self (Gordon-Roth, 2019; Locke, 2008, book 2, chapter 27).

9 Numerous views of selves see the ability for self-representation (understood in various ways) as a condition for selfhood (Dennett, 1988; Frankfurt, 1971; Locke, 2008). The condition of transparency can be understood in more or less robust ways, from rather "fictionalist" interpretations (Dennett, 1991; Metzinger, 2003), bordering on nihilism about selves, to more realist ones, which can correspond to different views of introspection (Byrne, 2012; Dretske, 1995; Shoemaker, 1996). Some deny that humans – prototypical selves – know themselves in a fully transparent way (Carruthers, 2011; Gopnik, 1993; Ryle, 1949), but they usually recognize that, even if the methods used to know oneself (and notably one's own mind) are not *fundamentally* different than the methods we use to know *others*, they are still applied in *importantly different ways*. I am inclined to believe that we are subjects to widespread introspective illusions, including regarding the very existence of phenomenal consciousness (Kammerer, 2016, 2019), which means I do not endorse transparency in any strong realist sense.

(c) *In order to be a self, a creature must have a certain degree of control over her behavior and her cognitive processes*.[10] A creature without any genuine control on her behavior and cognitive processes cannot be properly said to be a self.

I take these three features (coherence, transparency, control) to be *necessary* in order for a creature to be a self, but I do not claim that they are together *sufficient*: they do not constitute a *definition* of selves. One might want to *complete* this list with other necessary features : for example, the having of a wide enough variety of mental states, certain reasoning abilities, certain memory capacities, competences in social cognition, the capacity to grasp various kinds of normativity, etc. In fact, I find it *very plausible* that some (if not all) of these other features are also, at least to a certain degree, necessary in order for something to be a self.

I also take these three features to be rather *consensual* as necessary conditions for selfhood– at least for people who have a realist and naturalist conception of selves – which is why I do not intend to *argue* for them here. Moreover, arguing for this particular conception of selves is not the goal of my paper. I also take it that one can accept that these features are necessary for selfhood independently of the view one holds on the difficult issue of *diachronic personal identity* (Olson, 2019) – that is independently of any view on what grounds the identity of a particular self through time.

As I noted earlier, these three features come in degrees. When reflecting on how the property of "selfhood" itself depends on the degree to which these three features are present, two things seem to appear. First, it seems that, for these three features, we take it that there is a *threshold* corresponding to what is required for "proper selfhood" (adult humans being typically considered as "proper" selves). The exact definition of this threshold might be partially arbitrary (and thus debated), but it seems nevertheless to correspond to our practice: we think that a creature must have a certain degree of coherence, control and transparency in order to be a proper self.

Second, let's focus on the "downside" of the scale, that is, let's focus on creatures who instantiate these three features (as well as, perhaps, other necessary conditions for selfhood), but not enough to meet the threshold required for "proper selfhood". We usually do not want to say that these creatures – for example, infants, patients with severe dementia, intelligent non-human animals, etc. – are *selves* in the proper sense, but at the same time it seems too radical to entirely deny their selfhood. We are rather tempted to think of them as "proto-selves" or "diminished selves", or "quasi-selves" – or sometimes, using the vocabulary of persons: proto-persons, diminished-persons, quasi-persons, etc. (Ross, Ms). We thus express the idea that, even if an infant, an elephant, a patient with severe dementia, are not *proper selves*, they are still *closer* to proper selves than, say, rabbits or ants.[11]

These two facts suggest that *selfhood comes in degrees*, and is not an all-or-nothing feature. A creature can be *more or less* a self, and instantiate selfhood in a *more or less perfect way*. Adult humans, thus, probably instantiate selfhood more perfectly than eight year olds, who probably instantiate it more perfectly than young infants, patients with dementia, chimpanzees, elephants, etc. Moreover, these degrees of selfhood seem to depend at least partially on the degrees of instantiation of these three features: control, coherence and transparency (at least if we look at the downside of the scale).

Now, consider this. Let us admit that selfhood indeed comes in degrees. Let us also admit that the degree to which selfhood is instantiated correlates, on the downside of the scale (compared to adult humans) with the degree of instantiation of control, coherence and transparency – as suggested by the

10 A certain degree of control on behavior and cognitive processes is also accepted by numerous views as an essential condition for selfhood (Dennett, 1991; Ismael, 2016; Rorty, 1991; Ross, 2019).

11 Rabbits and ants might in turn be closer to selfhood than oysters or rocks (arguably, because they have some features, notably control, to a higher degree), but that is another matter. We usually do not want to apply the concepts of "self" or "persons" to them, even with qualification.

examples of infants, dementia patients, and highly intelligent animals such as chimpanzees or elephants. Let us finally admit that normal adult humans do not instantiate coherence, control and transparency to the *highest degree possible* – something I take to be *prima facie* extremely plausible. Adult humans do not instantiate coherence, control and transparency to the highest possible degree: metaphorically speaking, God could make a creature with more control, coherence and transparency than a normal adult human.

Now, what about such a hypothetical creature instantiating these three features more than a normal adult human? I think that we should probably think that this creature *also instantiates selfhood more* than a normal adult human. Indeed, if we have admitted that selfhood comes in degrees and correlates with these three features on the downside of the scale, it would be a remarkably lucky coincidence if this correlation suddenly broke down when we reached the exact level of coherence, control and transparency possessed by normal adult humans. Of course, this cannot be entirely and conclusively ruled out. However, it seems much more natural to suppose that such a creature would indeed be a *more perfect self* than normal adult humans.[12] Now, let us extend this reasoning, and think about a creature who would instantiate coherence, control and transparency to a *much higher degree* than normal adult humans – so that, for example, it would be as different from normal humans, regarding these features, as normal humans are different from young infants, dementia patients or chimpanzees. It is plausible that there could be such a creature. It then also seems plausible that this creature would thus instantiate *selfhood* in a *considerably* more complete and more perfect way than normal humans, so that they would be as different from humans, regarding selfhood, than humans are from diminished selves or proto-selves (say, infants, or even elephants). Let us call such an imaginary creature a "super-self".[13]

I will now consider that selfhood indeed comes in degrees, that it correlates with coherence, control and transparency, and that one can instantiate these features – and selfhood – *more* than the typical adult human (maybe to the point of becoming a *super-self*). On the basis of this conception of selves, let us turn to the two examples of technologies presented in the previous section. I will argue that these technologies can be seen as *self-building technologies*.

B/ iDiversity® and iFidelity® as self-building technologies

---

12 This does not at all imply that the typical adult human is *not* a proper self. Think, for example, of the property *being a democracy*. Certain countries are *more or less* democratic, which allows to distinguish between diminished democracies, endangered democracies, and *proper democracies*. But even countries which are prototypes of proper democracies in the contemporary world (Germany, France, the UK, the USA, etc.), arguably because they pass a (somewhat arbitrary) threshold regarding certain features (free press, free elections, respect of some basic rights, etc.), could be made *more democratic*. It might be that Switzerland, for example, is a *more perfect democracy* than Germany, France, the UK or the USA – which does not mean that these other countries are not "proper democracies" in some interesting sense. I suggest here that the same might very well be true of selves.

13 I think that the possibility of creatures instantiating selfhood *more* than normal humans, in virtue of them having more control, coherence and transparency (as to be *more perfect selves*) is more plausible than the possibility of creatures instantiating selfhood *much more* than normal humans, in virtue of them having much more control, coherence and transparency (as to be *super-selves*, as different from humans than humans are from diminished selves or proto-selves). Indeed: (1) While it is very plausible that normal adult humans have not reached the highest possible degree of coherence, control and transparency, whether or not they are *far away* from this highest possible degree is more of an open question. (2) Even if we admit that a creature endowed with *much more* coherence, control and transparency is possible, it is somewhat doubtful that this creature would also be proportionally more perfectly a self. Indeed, it is true that it would be a very lucky coincidence if the correlation between the degree to which these three features are possessed and selfhood, apparently observed on the "downside" of the scale, broke down *just when we reached the point of normal humans*. However, it would be somewhat less lucky (and thus less implausible) if the correlation broke down at some point which is *higher* on the scale, but nevertheless occurred *before* we reached the point of hypothetical super-selves. Hence, even though I think that the possibility of super-selves is plausible, I take it to be less plausible than the mere possibility of a creature instantiating selfhood somewhat more perfectly than typical normal humans (which I take to be *very* plausible).

This is a preprint version of a paper forthcoming in *AI & Society*. Please refer to the published version for citation purpose.

Let us go back to iDiversity® and iFidelity®. None of these imaginary technologies are supposedly *designed* to be a kind of self-building technology. The concept of self does not play any essential role in their conception, their design or their use. These technologies are arguably conceived and used to reach certain desired results which do not have much to do with selfhood: Pr. Truffle uses iDiversity® because he wants to avoid discriminating against black students, and Emma uses iFidelity® because she wants to avoid frustrations, angers, and have a happier marriage. However, I think that the functioning of these technologies makes it so that, as a matter of fact, were they to be implemented, they would constitute self-building technologies. Indeed:

a/ *These technologies improve the degree of* ***control*** *subjects have over their behavior and their cognitive processes*. Prior to the use of iDiversity®, Pr. Truffle arguably lacks control over many of his instinctive reactions when it comes to race-related issues: he wishes to think and to act in a non-racist way, but he does not. Emma, prior to the use of iFidelity, lacks control over some of her emotional reactions and some of her behavior: she wishes to act and feel like a loving, faithful wife, but she does not. iDiversity® enables Pr. Truffle to make his behavior and instinctive cognitive processes moreconform to what his deliberate and reflective intentions and beliefs are. It makes his behavior and his instinctive cognitive processes *less racist*, by suppressing the kind of input that triggers his racist behavioral and cognitive reactions. iFidelity® allows Emma to make her behavior and instinctive cognitive processes more conform to her deliberate intention, which is to be a loving, faithful wife. It considerably diminishes her fantasies about other men, as well as her frustration directed at her husband and the aggressive behavior that ensues, by suppressing the kind of input that triggers, in her, the problematic behavioral and cognitive reactions.

b/ *The technologies improve the degree of* ***psychological coherence*** *of the subjects*. Pr. Truffle, as well as Emma, before they start using the apps, can be described as having some sort of local psychological incoherence. There is a tension between, say, the *reflective and avowed* anti-racist beliefs and intentions of Pr. Truffle, and his *instinctive* reactions. Similarly, there is a tension between the reflective and avowed intention of Emma (to be a loving, faithful wife) and her *instinctive* emotional reactions (fantasizing about other men, being frustrated and angry with her husband)[14]. In both cases, technology reduces the incoherence. It modifies the manifested behavior and the instinctive cognitive and emotional reactions of the subjects (mostly by modifying the input they receive, so as to not trigger certain kinds of reactions) and make them *more in line* with their reflective and avowed beliefs and intentions.

c/ *The technologies improve the degree of* ***transparency*** *of the subjects' mental lives*. Prior to the use of the apps, the two subjects' mental life lacks a certain kind of transparency. When Pr. Truffle earnestly reflects by himself on whether or not he believes races are intellectually equal, he will 'sincerely' conclude that he does. It is only indirectly, by carefully observing his own behavior, that he comes to discover that he has instinctive racist reactions, at odds with his avowed and reflective beliefs. Similarly, when Emma (alone, at calm) earnestly reflects on whether or not she loves her husband and thinks of him as a wonderful, lovable person, she will 'sincerely' conclude that she does. And it is only indirectly,

14 There is an important philosophical debate about the correct way to describe the psychological tension at play in these cases. Take the case of Pr. Truffle. What does he *really believe*? Does he believe that races are equal, while his racist reactions do not really constitute *racist beliefs*, but simply racist behavioral dispositions ? Does he really believe that races are unequal, while he *pretends* to believe otherwise, and maybe *falsely believes* he believes otherwise? Does he hold two genuinely contradictory beliefs? Does he hold two contradictory beliefs, *but* in two different senses of beliefs? Does his case constitute an in-between case of belief? Eric Schwitzgebel (Schwitzgebel, 2010, p. 537) endorses this last interpretation in the case of Juliet, the implicit racist – from which the case of Pr. Truffle partially derives – and gives an overview of other possible interpretations. For other takes on similar cases, see (Frankish, 2016; Gendler, 2008; Hunter, 2011). These debates can be set aside now: the only thing that I need for my reasoning is the idea that there is a (local) deficit of psychological coherence in the case of Pr. Truffle (as well as in the case of Emma). I think that this idea is rather plausible, and that one could formulate it convincingly, whatever one's preferred option is when it comes to the correct precise formulation of the tension.

by observing carefully her own behavior, that she will come to discover that she has instinctive frustrated and angry reactions directed at her husband, at odds with her avowed and reflective beliefs and sentiments. In both cases, subjects fail to know transparently something important about their own mental lives, whether it is Pr. Truffle's implicit racism, or Emma's uncontrollable discontent with her husband, even though they might learn about these *opaquely*, by observing their own behavior. What is the effect of the technologies here? They suppress those of the subjects' *reactions* which are at odds with their avowed and reflective mental states – not by suppressing the *dispositions* to have these reactions to certain input, but by making it so (in a modally robust way) that the relevant input just never obtains. Therefore, there is a sense in which these technologies *reduce* the part of the subjects' mental lives which is at odds with the way in which the subjects transparently self-represent when they reflect on themselves – even if the *dispositions* to have the problematic reactions remain (as inactivated dispositions). Therefore, the technologies improve the degree of transparency of the subjects' mental lives. Not because they make the subjects' capacity of self-representation *more effective*, but because they modify some aspects of the subjects' mental lives (their instinctive reactions) and make them *more similar* to these aspects of the subjects' mental lives (avowed and reflective mental states) which are transparently and correctly grasped by their self-representation.[15]

To summarize, these two technologies improve the psychological coherence, transparency of mental life, and control of behavior and cognitive processes of the subjects who use them. Control, transparency and coherence are, as I said, *crucial features* of selfhood, which *come in degrees* and which correlate with the degree of instantiation of selfhood. Although this conclusion could be resisted in different ways (see the next section), one can make the case that it raises the degree to which the corresponding subjects satisfy selfhood: iDiversity® and iFidelity® make (locally) their users *more perfect selves*. They are *self-building technologies*.

C/ Future self-building technologies

If my argument is correct, self-building technologies are already possible, given the current state of technology (more or less). What about possible *future* self-building technologies – made possible by *future* technological progress? How would it depart from the two examples I just analyzed?

There are three relevant factors that we should take into account when we try to speculate about these even more distant possible technologies: *design*, *performance* and *integration*. These three factors could differentiate these distant possible technologies from the two imaginary technologies I described earlier, and consequently make them relevantly different when it comes to their self-building features.

First, *design*. The two imaginary technologies I described are not *intended* to be self-building technologies: they are simply not designed as such. They are hypothetically designed to achieve certain psychologically and socially relevant goals: avoid racial discrimination, contribute to a happy marriage. However, it is not implausible that, at some point in a more distant future, we could see the rise of self-building technologies which are *designed as such* (with the very goal of perfecting selfhood in mind). The improvement of control, transparency and coherence which partly constitute selfhood would not just be a *means* to achieve these other socially or psychologically relevant goals, or a by-product of the achievement of these goals, but the very goal pursued by such technologies. We can expect the said improvement to be thus substantively more important, and more comprehensive.

Second, *performance*. We can expect future AI technologies to be more powerful and efficient than the ones which are available now. If we think about a *distant future*, we can imagine that the software that will be available, compared to iFidelity® or iDiversity®, will be much more effective .

15 This effect of these two technologies can also be seen as an improvement of introspection, if one has a liberal conception of introspection according to which this kind of technologically-mediated self-shaping can count as introspection. For introspection as self-shaping, see (Schwitzgebel, 2014, sect. 2.3.2).

They will be able to process a larger range of inputs, detect more complex and more subtle patterns, modify these patterns in subtler and more flexible ways, following more complex rules and aiming at more complex goals.

Third, *integration*. iFidelity® and iDiversity® are only loosely cognitively and functionally integrated with Emma and Pr. Truffle. Emma can easily take off her glasses, or deactivate the app; Pr. Truffle can easily decide to switch to another way to communicate with his students (personal email, say), or make attempts at identifying them if he really wants to. Moreover, in both cases, the apps operate by modifying the *input* received by Emma and Pr. Truffle (and, for iDiversity®, by making suggestion for modifying the written output of the professor), but do not have a direct impact on the cognitive processes of the subject, which remains more or less untouched. On the other hand, we could imagine that the self-building technologies available in a distant future could be much more tightly integrated to the subjects: they could take the form of tightly integrated "cognitive modules" (Bostrom & Sandberg, 2009, p. 320-321; Schneider, 2009, 2019) which would directly affect the details of the perceptual, emotional, cognitive processes of subjects. Moreover, in the scenario in which future humans decide to completely "merge with AI" and proceed to *cognitive uploading* (Kurzweil, 2006; Oxford University, 2008; Schneider, 2019), the self-building apps could consist, not only in *additions* to the normal cognitive functioning of the subjects, but in deep modifications of their cognitive architecture.[16]

Because of these three factors – design, performance, integration – I think we can expect the self-building technologies that technological progress could provide in a distant future to be much more efficient at raising the control, transparency and coherence of future subjects – in ways that might be hard to fully imagine or comprehend at our stage.

I argued earlier that iFidelity® and iDiversity® were potential *self-building* technologies: that they could make their users more perfect selves. However, the difference between Emma and Pr. Truffle *prior* to the use of the apps and *after* the use of the apps arguably remains a small difference of *degrees* when it comes to the instantiation of selfhood. It appears similar to the kind of difference we draw between two "normal" adult humans who differ substantively (but not radically) when it comes to control, transparency or psychological coherence – for example, because one, but not the other, is what we would call a "disciplined", self-reflective", "reliable" and "tempered" individual. On the other hand, we can imagine that the difference in terms of control, transparency and coherence, between a normal adult human and a subject using one of these possible self-building technologies made possible in a distant future, could be much more striking. Such difference would perhaps be closer to the difference between normal adult humans and diminished selves, proto-selves or almost-selves (elephants, apes, young infants, severe dementia patients, etc.), and would perhaps call for the use of a different term.

16 The kind of cognitive integration that is required for such modules or modifications to be genuinely a *part* of the *mind* of the subjects depends on the kind of view one holds regarding the "extended mind thesis" (Clark & Chalmers, 1998). This might then affect whether or not one sees the corresponding technologies as genuine self-building technologies or not – for example (as will be discussed in the next section) one might require *selfhood* to be only grounded in the psychological features and capacities that depend on the *mind* of the subject, and not on the mind-plus-its-environment (including technological artifacts).

Such putative subjects could be said to be "super-selves".[17] This would hold independently, I think, of whether or not they can be said to be *super-intelligent (Bostrom, 2016)*.[18]

Let us take stock. I presented two possible technologies, which could be (more or less) implemented given our current state of technological progress. I argued that these technologies should be considered *self-building technologies*. I also argued that future technological progress could allow for more disrupting and radical self-building technologies, able to give rise to *super-selves* – cognitive systems instantiating selfhood way beyond what "normal" adult humans do.

My theses contrast with two common stream of thoughts when it comes to the potential effects of AI technology on selfhood. First, it contrasts with a form of *pessimism* regarding AI technology – the idea that using AI for cognitive enhancement might lead to a loss of selfhood, if not to a complete destruction of our selves (Agar, 2010; Schneider, 2019) – through the loss of consciousness or the disruption of trans-temporal personal identity. Second, it also contrasts (although for different reasons) with the idea that using AI for cognitive enhancement might lead to increased *inter-individual* cognitive integration, and then to a loss of selfhood (at least at the level of the human individual). Future humans would then form some kind of collective *hive mind* (a common theme in science-fiction since (Stapledon, 1930); see also (Kammerer, 2015; Sandberg, 2003; Schwitzgebel, 2015)), within which "selves" would not exist anymore (or at different degrees and/or levels). In contrast, my reasoning stresses the potential use of technology to *increase selfhood* at the level of human individuals: we could use AI technology for enhancement, in a way that would make us more genuine and more perfect selves.

It is worth noting, though, that I speak simply here of a *contrast* between my view and these theses: I do not claim here that my view *refutes* or even *contradicts* "pessimistic" predictions regarding the impact of technological progress on future selves. It might be, after all, that self-building technologies will be available, but that they will not be used, or not widely; or that their use will not counterbalance stronger forces, leading to a loss of selfhood (through a destruction of selves, or through their integration in a collective hive mind). However, I also think that the discussion regarding the

---

17 In this paper I use the term of "self" rather than the term "person", partly because "person" has stronger normative connotations – and I want to avoid discussing normative issues here. However, one can legitimately wonder whether such "super-selves" would have, in virtue of their super-selfhood, some extra rights (and/or duties) compared to us "standard" selves – the same way it seems that our selfhood gives us some extra rights (and duties) compared to proto-selves, diminished selves or quasi-selves. I do not intend to give an answer to this question here, although I think that, if super-selves have extra rights and duties, they are likely to be linked to their extra-capacities. To give an example of what I have in mind: maybe super-selves would have, compared to "standard" selves, a much more stringent and absolute *right* to receive only precise, accurate and correct information, given that the higher degree of control possessed by extra-selves also means that their forming a reflective belief or desire (possibly on the basis of the information they are fed with) has a much more long-lasting causal impact. If I convince a *standard self* of an incorrect ethical view (say, because I gave them false information, out of mere sloppiness), I certainly did something bad. However, if I thus convince a super-self, I probably committed a much worse crime, given that this super-self has a far superior capacity to *enforce* this view and to modify robustly its future behavior in accordance with the view (while most of the behavior of a "standard self" will probably be generated anyway by a mix of habits, innate emotions and desires, intuitions, in a way that is partially independent of their reflective ethical beliefs).

18 I am not claiming here that there is simply *no relation* between intelligence, or various putative aspects of general intelligence (inferential ability, working memory capacity, etc.) and the degree of instantiation of selfhood. However, it is worth noting that, at least on paper, the two are relatively independent. We could easily imagine an extremely intelligent human being (at least on some standard meaning of "intelligence") who would be a *less perfect self* than the average adult human, because they would lack coherence, control or transparency (think about the caricature of the "mad genius", who might be incredibly rare in reality but nevertheless seems like a possibility). On the other hand, we could easily imagine someone who is below average when it comes to intelligence, but shows control, coherence and transparency to a very high degree, and therefore instantiates selfhood more perfectly than most (including most intelligent people).

potential and/or predictable effects of technological progress on our selves would benefit from paying more attention to potential self-building technologies, such as the ones I described.

3. Objections

My view on self-building technologies can be subjected to numerous objections. I will now examine three of them.

(A) "There can be no self-building technologies, because there can be no selves, or because selves are primitive entities that cannot be built, or because selfhood is an all-or-nothing feature which does not come in degrees."

It is true that my arguments presuppose the falsity of *nihilism* regarding selves.[19] It also presupposes that selves are not *primitive* entities, but that the selfhood of selves is constituted by a certain kind of functioning of the creatures who count as selves. Finally, my view does indeed suppose that selfhood can, in an interesting sense, come in degrees (even though nothing prevents us from determining useful thresholds, for example for "proper selfhood"). However, I think that even someone who denies one of these presuppositions, and then rules out the possibility of self-building technologies, can reinterpret my argument charitably so as to make the truth-value of the thesis I defend an *open question*. For example, let us say that, instead of *selfhood* (which is either uninstantiated, or primitive, or does not come in degree), one focuses on selfhood*. Selfhood* is a real, natural, composed property, which comes in degrees, and in virtue of which the creatures that we usually call "selves" function in ways which make it so that *we can successfully treat them like selves* – notably because they have enough control, transparency and coherence to respond appropriately to at least *some* of the expectations we associate with the term "self". One can then read my argument as an argument bearing on the possibility of *self*-building technologies*. The thesis might then have different implications and connotations, but it might still lead to ask open and interesting questions.

(B) "What you describe as self-building technologies do not at all "build" selves! What they do is much more mundane and ordinary: they are nothing but imaginary devices for self-blinding or self-nudging. At best, all these technologies can do is help fight weakness of the will, or implicit biases; but they do not *build selves.*"

I agree that these are possible ways to describe what iFidelity® and iDiversy® do. Both technologies consist in *self-blinding* (to male beauty for Emma, to the race of his students for Pr. Truffle); iDiversity® also has aspects of *self-nudging* (as the software makes suggestions regarding what Pr. Truffle writes). Both technologies can be seen as ways of fighting weakness of the will (without the app, Emma cannot help looking at beautiful men and admiring them; Pr. Truffle cannot help being biased against his black students), and in the case of iDiversity®, fighting implicit biases. However, describing these technologies in such a way is not necessarily in opposition with a description of them as self-building technologies. Seeing them as self-building technologies is just *another way* to look at them, which I think is interesting and relevant (compare: a new technology to *build rockets* could also be seen at the same time, more mundanely, *a mere new way to put together metal, plastic, electronic components etc.* Both descriptions could be true at the same time, although one here is maybe more useful and telling than the other).[20]

---

19 At least if such nihilism comes with some modal force, and states that *there can be no selves* given, say, the current laws of nature. There is a possible nihilist view about selves that says that *there are currently no selves* but that *there might be some,* given our current laws of nature. One might then wonder whether or not self-building technologies of the kind I presented would be capable of "creating" such selves. I will not explore this particular position (for which I have sympathy) any more here.

20 Moreover, it might be interesting to recognize that, even if these technologies can be seen as self-nudging or self-blinding technologies (or technologies used to fight weakness of the will or implicit biases), they do all of that

(C) "What you describe as self-building technologies do not really "build" selves! Indeed, these technologies merely amount to *changing the environment* of the subjects, so that they no longer perceive certain things that make them act in ways they dislike. Pr. Truffle does not *build his self* more by using iDiversity® than by moving to a state in which there are virtually no black people (say, Montana). Emma does not *build her self* more by using iFidelity® than if she simply went to live forever on her remote vacation village in Scotland. In all these cases, there is no change of the individual itself (but only a change in the external environment), and therefore no genuine self-improvement."

I think this is quite a serious concern. To answer it, first of all, I suggest we set aside the moral and normative connotations of the expression "self-improvement". The question I want to consider here is not whether or not what Emma and Pr. Truffle do is *praiseworthy*, or whether or not it is *less* praiseworthy than if they manage to become less racist or more sentimentally constant by other, "natural" means.[21] The question here is only whether or not these two devices can be said to make them more perfect selves (by raising their degree of control, coherence, transparency). Maybe there are ways to become *more perfect selves* which are not praiseworthy (e.g. think about an imaginary "selfhood pill" that we could give to apes or elephants, to give them more control, and make them more coherent and transparent. There would be nothing praiseworthy about the apes taking the pill, but they would still become more perfect selves).

Now, the core concern remains: aren't these two technologies nothing more than sophisticated ways of changing the *environment* of the subjects? One can begin by noticing that, even if it is the case, these technologies consist in changing, in a systematic and counterfactually robust way, an extremely localized part of the *proximal* environment of the subjects (what is displayed on Pr. Truffle's screen, what goes through Emma's glasses). This is very different from "classical" changes of the environment, which require much more modification. This kind of subtle, localized and robust change of the environment in turn allows the subjects to reach goals that they could not attain otherwise: Pr. Truffle arguably wants *to interact with black students in a fair way* (while, moving to Montana, he could at best merely *cease to interact with black students in an unfair way*). Similarly, Emma wants to be able *to interact with the people she ordinarily interacts with, without having unfaithful trains of thoughts and emotions* (which is not something she could do if she had to move out to a remote Scottish village).

Part of the concern, again, remains. At this point, the defender of the possibility of self-building technologies has three options. The most radical would be to deny that these two technologies simply consist in *external devices* able to change (an extremely localized part of) the proximal environment of the subjects: in fact, they can genuinely count as *a part of the subjects themselves, as they are part of their minds or cognitive systems*. Someone who is attracted to a radically extended conception of the mind might find this answer satisfying (Clark, 2008; Clark & Chalmers, 1998; Hutto & Erik, 2013). In this conception, iFidelity® and iDiversy® are simply part of the (extended) minds of the subjects. Given that this answer implies that they are not mere *external* devices (changing the proximal environment of the subject), it becomes easier to defend the view that they really contribute to building the selfhood of the subjects. Another, slightly more concessive version of the same answer, would grant that the two

---

in substantively different ways than our traditional techniques and technologies. For example, our "traditional" ways of fighting weakness of the will (through personal or institutional commitment, for example) usually impact only some limited (though decisive) actions (to which we commit). iDiversity® and iFidelity®, on the other hand, "put the will" of the subjects where the will usually is never able to go. It allows the reflective and avowed beliefs and desires of the concerned subjects to systematically shape the details of some of their cognitive and emotional processes (how Pr. Truffle thinks and feels when he communicates with black students, what Emma thinks and feels when she interacts with attractive men).

21 I do not take these concerns, which are related to the more general concern that AI technology might have a "deskilling effect" (Vallor, 2015) in the moral domain, to be baseless or uninteresting. However, I take it that they fall beyond the scope of this paper.

technologies I described cannot count as genuine self-building technologies, because they are not integrated enough, from a functional point of view, to be a genuine part of the extended mind of the subjects. However, *similar* technologies with a higher degree of functional integration within the cognitive systems of Pr. Truffle and Emma *would*. Maybe, for example, Emma simply *changes* (systematically) her environment whenever she wears glasses and uses iFidelity®. However, she would not only change her *environment* but also her *mind* (and thus build her *self*) if, instead of glasses, she was wearing lenses (surgically implanted and effortful to reprogram, say) doing the same job. (This is just an example: the exact degree and nature of cognitive integration required for the device to start modifying the *self* or the *cognitive subject itself*, rather than its environment, will depend on the exact stance one takes on the issue of *extended mind*). This concessive variation admits that the two cases I described are not genuine cases of self-building technologies, but that they might still be relevant to the argumentation in favor of the possibilities of such technologies, as they indirectly suggest that self-building technologies could be implemented in the form of more tightly integrated analogous devices.

The second available answer amounts to admitting that the two technologies I describe indeed consist in mere external devices which simply provide (systematic and counterfactually robust) ways to change (an extremely localized part of) the proximal environment of the subjects. However, one could then add that it is not true that the *self* can only be constituted by features of the *mind* of the subjects. In this view, even if one denies that the two described devices are part of the *extended mind* of the subjects (maybe because there is no such thing as an "*extended mind*"), they can still be part of their "*extended self*". An argument for the possibility of such an extended self could perhaps be built by appealing to other, more mundane examples in which it seems plausible that some objects or features *external* to the *minds* of subjects (belonging to their artifactual, perceptual, linguistic, social and technological environment) nevertheless contribute to constitute them as *selves*. Wouldn't I lose (not only causally, but constitutively) some of my coherence, control and transparency (and thus some of my selfhood) if I was suddenly put in complete isolation, in a sensory deprivation tank, without any ability to communicate or interact fruitfully with my usual environment?

The third possible answer grants to the opponents that such technologies cannot be genuinely self-building – for example, if one denies that there can be any form of *extended mind* or *extended self*, and that selfhood must be grounded in biological features and capacities, the functioning of which occurs say, inside the human body. However, similar to my previous response to objections inspired by primitivist or nihilist conceptions of selves, I then recommend to consider the concept, not of *selfhood*, but of *selfhood***, which is a property which can be grounded both in features of biological individuals *and* in features of their environment. Selfhood** is the property in virtue of which the creatures that we usually call "selves" function in ways which make it so that *we can successfully treat them like selves*. One could then deny that the technologies I describe count as self-building technologies, by nevertheless considering that they might count as self**-building technologies. This thesis, again, might then have different implications and connotations, but it might nevertheless be important to discuss it.

(D) "There is nothing really new in the effect of the technologies you describe, or in the idea of self-building technologies. We *already* build selves, using various devices (and some, though not all of them, imply the use of technology). Various existing widespread practices – giving and using permanent proper names to refer to individual humans, striving to tell coherent life-stories about ourselves, committing to others by formal and informal means, confessing and regretting our faults, being punished for our crimes, writing curriculum vitae or diaries, collecting memories of episodes of our lives through photos, texts, recordings, etc. – can already be meaningfully interpreted as *self-building* practices."

There is a sense in which all of that is entirely true – but I do not see it as a problem for my view. The self-building technologies I described earlier in this paper are not the *only things* that might contribute to building selves. To a great extent, one might say that the selves that we *already are* have been built (and are continuously built) by various practices. Some of them are widespread social

practices (giving unique and permanent proper names to the "same" individuals (Bourdieu, 2004) and expecting them to answer to these names (Althusser, 1970), surveilling and punishing individuals for "their" faults or crimes, etc. (Foucault, 1995). Some are less widespread, and require continuous and deliberate efforts of the subjects themselves: they are often studied, from a philosophical and historical perspective, under the expression "techniques of the self" (Foucault, 1988, 1990; Hadot, 1995, 2002). I think that one can see the hypothetical "self-building technologies" I described earlier as an extension, with different technological means, of these older "techniques of the self".[22] What is interesting to note, of course, is the considerable increase in means offered by technology. AI-based self-building technologies allow subjects to do things that none of the traditional "techniques of the self" could do[23]: correspondingly, their effects, when it comes to self-building, can be substantively more important.

4. Reservations

I defended the thesis that self-building technologies are possible, given (more or less) our current state of technological progress. I also argued that future technological progress might allow for some more radical self-building technologies – which could give rise to *super-selves*.

Now, setting aside the *objections* that one could make to these theses, I think that one might also have a number of more general *reservations* regarding self-building technologies. I will now examine two of them.

(A) "You described some imaginary technologies which could raise the degree of control, transparency and coherence of individuals, and make them *more perfect selves*, which sounds very positive. However, very similar technologies, if controlled by a totalitarian state, by powerful and corrupt capitalistic companies, or other nefarious agents, could lead to the most frightening and inescapable situations of *subjection* of individuals. Far from *building* selves, these technologies would *enslave* them – and it is not even clear that these technologically enslaved selves would still count as selves. Describing these technologies as "self-building" might indirectly obscure the potentially terrifying effects they might have, and motivate a dangerous, irrational form of techno-optimism".

I share, to a great extent, this reservation. The technologies I described are technologies in which the intervention of a (rather simple) AI software allows to robustly *change* the perceptual input received by a subject, in a systematic way. I describe how such a device could be used by an individual to increase their own control, coherence and transparency. However, it is also clear that, if *another agent* controls

---

22 I suggested here that we could interpret self-building technologies as an extension of a set of social practices and techniques. Another relevant way to interpret such technologies is to see them as pursuing a kind of hierarchical process of self-building which is already at play on a *biological* level, and independently in partof culturally-dependent social practices. It has been argued, for example, that our minds have a two-layered structure – of *minds* and *superminds* (Frankish, 2004). Self-building technologies could be seen as a way to reinforce or to extend the biological *supermind*, and maybe to create a new supermind which would consist in a more or less integrated artifactual-cum-biological system.

23 Just think of the considerable infrastructure that would be required to produce the equivalent of iDiversity® without the use of technology. One would need to hire people whose jobs would be to systematically anonymize and "racially neutralize" all communications between students and a professor, and make rewriting suggestions. Although it would be in theory possible to do so, the huge cost would make such a pre-technological solution impracticable in the long run in the context of students/professor interactions. Of course, *similar* infrastructure already exists at a supra-personal level – companies and institution hire *diversity managers*, whose jobs are partly to do precisely what iDiversity® does for Pr. Truffle. However, because such diversity managers usually do not work at the scale of the individual, their work cannot be said to have *self-building effects* (although it arguably has *socially desirable* effects). Similarly, think of the considerable cost that Emma would need to pay in order to be able to interact with attractive men without having to look at them, without the help of technology. The possibilities that come to mind – having an army of employees actively hiding any handsome man who is around, say, or literally blinding herself – would anyway probably prevent her from living the "normal" life she aspires to.

this device, this agent could gain considerable power over the individual whose perceptual access to the world is thus modified. Such an agent could literally determine *how the individual sees the world*, which comes a long way to controlling how they think, want and act. Eric Schwitzgebel's story ("My daughter's rented eyes"), which I cited previously, is a fascinating description of a possible situation in which control of the perceptual (visual) input of a growing part of the population can lead to a form of inescapable subjection of individuals, which makes Orwellian worlds pale in comparison.

Whether the technologies I described serve as self-building technologies, or as domination and subjection technologies, essentially depends on *who* controls them. Are they under the control of the subjects themselves – that is, does their functioning directly and exactly depend on the avowed and reflective beliefs and intentions of the subjects who will then have their own perceptual input modified? Or are they under the control of *other* agents – whether these agents are personal (other individuals) or supra-personal (states, companies, etc.)? Note here that, in this respect, the case of these technologies bears some similarity to the case of the *techniques of the self* that I mentioned earlier. The various pre-technological social practices which can be used to build selves *by individuals themselves* (Foucault, 1988, 1990; Hadot, 1995) can be compared with neighboring techniques which serve primarily for social control (Althusser, 1970; Foucault, 1995). It is also worth noting that, when I say that whether or not these technologies end up building selves or enslaving them depends on who controls them, I do not simply mean here "who *decides* to use them". It is easy to imagine a situation in which the *decisions* to use technologies of this kind are made by the individuals which will be impacted by them (so that these technologies cannot start modifying their perceptual input without their prior informed consent), while the overall effect of such technologies is nevertheless subjection, and not self-building. Schwitzgebel's story gives a powerful imaginary example of such situations: individuals *consent* to use the artificial eyes provided by the Eye & Ear Company, because the benefits of the artificial eyes are such that refusing to use them would bear too much of a cost for each of them. However, once they have consented, they simply do not have control over the details of the ways in which their perception of the world is modified – and the company acquires a decisive power over the way in which they perceive the world. Therefore, what is crucial in order to avoid similar technologies to lead to subjection, is for the individual not only to *have control over whether or not to use them*, but also to have *continuous* control over *how exactly* these technologies modify their perceptual input. What matters is not merely to implement the respect of *consent*, or even *informed consent* of individuals; what matters is the *control* and the *effective power* that individuals have over these technologies. Individuals must be able to decide reflectively when and how to use these technologies, how exactly they function, which input they modify, in which situations, according to which rules, etc. – and only then can these technologies be likely to have primarily self-building effects. In the end, this kind of control cannot obtain outside of a wider social context in which individuals have more generally effective power over their lives and their environment. Individuals who are granted entire and permanent control on (potentially) self-building technologies, but who are at the same time subjected to extremely strict constraints of efficiency or profitability (enforced by the state or generated by the overall economic organization) in order simply to sustain their own biological or social existence, will be unlikely to "build their own selves" in any relevant sense.

(B) "Let us admit that the technologies you describe are indeed controlled directly by individuals, in a way that makes them genuine *self-building* technologies. Does that really mean that the use of such technologies is desirable? Do we really want to be more perfect selves? *Should* we want that? After all, increasing our degree of control, transparency and coherence might also be seen as a way to *reduce* our internal complexity, spontaneity and plurality. It might be that some of the value of our lives comes from us *not being perfect selves* – from us lacking a certain kind of coherence, transparency and control (because it has a certain kind of *intrinsic value*, or because it has *instrumental value*: it makes us more adaptable, or more creative, etc.)"

I share these reservations regarding self-building technologies. There is a philosophical stream of thought which emphasizes the value of being, to a certain extent, spontaneous, uncontrolled, divided and even *opaque* to oneself (Kammerer, 2009; Nietzsche, 1974, §143, 1992). Whether or not becoming more perfect selves is something that we should strive for is, in my mind, an open question – and a question certainly worth asking. It could seem that I implicitly presupposed, in this article, that self-building technologies are a *good thing*, and that we *should* become more perfect selves, but I made no such presupposition. I take it to be true, though, that many of us *would like* to be more perfect selves – increase their degree of control, coherence and transparency. I think that, by arguing for the *possibility* of self-building technologies, I have also given more reasons to think hard about this question: should we really try to become more perfect selves, to what extent, and in which ways? What would we win by doing so? What would we lose?

5. Conclusion

I argued that self-building technologies are possible, given the current state of technology, and that future technological progress might provide us with *radical* self-building technologies, able to transform us into *super-selves* – as different, maybe, from "normal" selves, than "normal" selves are from diminished selves or proto-selves. This possibility should make us deeply and urgently concerned both about the possible use of these technologies – which could easily be recruited, not to *build* selves, but to subject and dominate them – and about the *value* of instantiating selfhood in a more perfect way.

References

Agar, N. (2010). *Humanity's End: Why We Should Reject Radical Enhancement*. https://doi.org/10.7551/mitpress/9780262014625.001.0001

Agar, N. (2012). On the irrationality of mind-uploading: a rely to Neil Levy. *AI & SOCIETY*, *27*(4), 431-436. https://doi.org/10.1007/s00146-011-0333-7

Agar, N. (2014). On the Prudential Irrationality of Mind Uploading. In *Intelligence Unbound* (p. 146-160). https://doi.org/10.1002/9781118736302.ch9

Althusser, L. (1970). Ideology and Ideological State Apparatuses (Notes towards an Investigation). In *Lenin and Philosophy and Other Essays*. Verso.

Bostrom, N. (2016). *Superintelligence: Paths, Dangers, Strategies* (Reprint). Oxford, United Kingdom ; New York, NY: OUP Oxford.

Bostrom, N., & Sandberg, A. (2009). Cognitive enhancement: methods, ethics, regulatory challenges. *Science and Engineering Ethics*, *15*(3), 311-341. https://doi.org/10.1007/s11948-009-9142-5

Bourdieu, P. (2004). The Biographical Illusion. In P. DuGay, J. Evans, & P. Redman (Éd.), & Y. Winkin & W. Leeds-Hurwitz (Trad.), *Identity: A Reader* (p. 297-303). London: Sage Publications.

Byrne, A. (2012). Knowing What I See. In D. Smithies & D. Stoljar (Éd.), *Introspection and Consciousness* (p. 183-210). New York: Oxford University Press.

Carruthers, P. (2011). *The Opacity of Mind*. Oxford: Oxford University Press.

Chiang, T. (2002). *Stories of Your Life and Others*. Tor Books.

Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. New York: Oxford University Press.

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

Dennett, D. (1988). Conditions of Personhood. In M. Goodman (Éd.), *What Is a Person?* (p. 145-167). Clifton, New Jersey: Humana Press.

Dennett, D. (1991). *Consciousness Explained*. Penguin.

Dretske, F. (1995). *Naturalizing the Mind*. MIT Press.

Foucault, M. (1988). *Technologies of the Self: A Seminar with Michel Foucault* (1st edition; L. H. Martin, H. Gutman, & P. H. Hutton, Éd.). Amherst: University of Massachusetts Press.

Foucault, M. (1990). *The History of Sexuality, vol 2: The Use of Pleasure* (R. Hurley, Trad.). New York: Vintage Books.

Foucault, M. (1995). *Discipline and Punish: The Birth of the Prison* (2nd Vintage Books ed). New York: Vintage.

Frankfurt, H. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, *68*(1), 5-20.

Frankish, K. (2004). *Mind and Supermind*. Cambridge: Cambridge University Press.

Frankish, K. (2016). Playing Double. In M. Brownstein & J. Saul (Éd.), *Implicit Bias and Philosophy, Volume 1* (p. 23-46). https://doi.org/10.1093/acprof:oso/9780198713241.003.0002

Gendler, T. S. (2008). Alief and Belief. *The Journal of Philosophy*, *105*(10), 634-663. Consulté à l'adresse JSTOR.

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, *16*, 1-14.

Gordon-Roth, J. (2019). Locke on Personal Identity. In E. N. Zalta (Éd.), *The Stanford Encyclopedia of Philosophy* (Spring 2019). Consulté à l'adresse https://plato.stanford.edu/archives/spr2019/entries/locke-personal-identity/

Güera, D., & Delp, E. J. (2018). Deepfake Video Detection Using Recurrent Neural Networks. *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1-6. https://doi.org/10.1109/AVSS.2018.8639163

Hadot, P. (1995). *Philosophy as a Way of Life: Spiritual Exercises from Socrates to Foucault* (A. Davidson, Éd.). Malden, MA: Wiley-Blackwell.

Hadot, P. (2002). *What is Ancient Philosophy?* (M. Chase, Trad.). Cambridge (Mass.): Harvard University Press.

Hunter, D. (2011). Alienated Belief. *Dialectica*, *65*(2), 221-240. Consulté à l'adresse JSTOR.

Hutto, D., & Erik, M. (2013). *Radicalizing Enactivism: Basic Minds Without Content*. Cambridge, MA: MIT Press.

Ismael, J. (2016). *How Physics Makes Us Free*. Oxford, New York: Oxford University Press.

Kammerer, F. (2009). *Nietzsche, le sujet, la subjectivation. Une lecture d'Ecce Homo*. Paris: L'Harmattan.

Kammerer, F. (2015). How a Materialist Can Deny That the United States is Probably Conscious - Response to Schwitzgebel. *Philosophia*, *43*(4), 1047-1057.

Kammerer, F. (2016). The hardest aspect of the illusion problem - and how to solve it. *Journal of Consciousness Studies*, *23*(11-12), 123-139.

Kammerer, F. (2019). The illusion of conscious experience. *Synthese*. https://doi.org/10.1007/s11229-018-02071-y

Kurzweil, R. (2006). *The Singularity Is Near: When Humans Transcend Biology*. New York: Penguin Books.

Levy, N. (2011). Searle's wager. *AI & SOCIETY*, *26*(4), 363-369. https://doi.org/10.1007/s00146-011-0317-7

Locke, J. (2008). *An Essay concerning Human Understanding*. Oxford, New York: Oxford University Press.

Metzinger, T. (2003). *Being no one*. Cambridge (Mass.): MIT Press.

Nietzsche, F. (1974). *The Gay Science: With a Prelude in Rhymes and an Appendix of Songs* (1 edition; W. Kaufmann, Trad.). New York: Vintage.

Nietzsche, F. (1992). *Ecce Homo: How One Becomes What One Is; Revised Edition* (Reprint edition; M. Tanner, Éd.; R. J. Hollingdale, Trad.). London, England ; New York, N.Y: Penguin Classics.

Olson, E. T. (2019). Personal Identity. In E. N. Zalta (Éd.), *The Stanford Encyclopedia of Philosophy* (Fall 2019). Consulté à l'adresse https://plato.stanford.edu/archives/fall2019/entries/identity-personal/

Oxford University. (2008). *Whole Brain Emulation: A Roadmap.* (Technical Report Nº #2008-3). Future of Humanity Institute.

Rorty, A. O. (1991). *Mind in Action: Essays in the Philosophy of Mind* (Reprint). Beacon Press.

Ross, D. (Ms). Addicts and elephants: two varieties of diminished persons. *Manuscript*. Consulté à l'adresse https://www.google.fr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=2ahUKEwjTofvAtb7kAhWRPFAKHac9D0YQFjAAegQIBRAB&url=https%3A%2F%2Fwww.academia.edu%2F31618031%2FAddicts_and_elephants_two_varieties_of_diminished_persons&usg=AOvVaw1X_vU6Ym9mRkjiPJuSyNFi

Ross, D. (2019). Consciousness, Language, and the Possibility of Non-human Personhood: Reflections on Elephants. *Journal of Consciousness Studies*, *26*(3-4), 227-251.

Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson.

Sandberg, A. (2003). We, Borg. Speculations on Hive Minds as a Posthuman State. Consulté 17 septembre 2019, à l'adresse https://www.aleph.se/Trans/Global/Posthumanity/WeBorg.html

Schneider, S. (2009). Future Minds: Transhumanism, Cognitive Enhancement, and the Nature of Persons. In V. Ravitsky, A. Fiester, & A. Caplan (Éd.), *The Penn Center Guide to Bioethics*. Springer.

Schneider, S. (2019). *Artificial You: AI and the Future of Your Mind*. Princeton University Press.

Schneider, S., & Mandik, P. (2018). How Philosophy of Mind Can Shape the Future. In A. Kind (Éd.), *Philosophy of Mind in the Twentieth and Twenty-first Centuries* (p. 303-319). New York: Routledge.

Schwitzgebel, E. (2010). Acting contrary to our professed belief or the gulf between occurent judgment and dispositional belief. *Pacific Philosophical Quarterly*, *91*, 531-553.

Schwitzgebel, E. (2014). Introspection. In E. Zalta (Éd.), *Stanford Encyclopedia of Philosophy (Summer 2012 Edition)* (Summer 2014 Edition). Consulté à l'adresse http://plato.stanford.edu/archives/sum2014/entries/introspection/

Schwitzgebel, E. (2015). If Materialism is True, the United States Is Probably Conscious. *Philosophical Studies*, *172*(7), 1697-1721. https://doi.org/10.1007/s11098-014-0387-8

Schwitzgebel, E. (2019). *A Theory of Jerks and Other Philosophical Misadventures*. MIT Press.

Shoemaker, S. (1996). *The First-Person Perspective and Other Essays*. Cambridge University Press.

Stapledon, O. (1930). *Last and First Men: A Story of the Near and Far Future*. London: Methuen.

Vallor, S. (2015). Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character. *Philosophy & Technology*, *28*(1), 107-124. https://doi.org/10.1007/s13347-014-0156-9