Ethical Consequences of Autonomous AI. Challenges for Empiricist and Rationalist Philosophy of Mind

Patrizio Lo Presti[†]
patrizio.lo_presti@fil.lu.se

ABSTRACT

The possibility of autonomous artificially intelligent systems (AAIs) has awaken a well-known worry in the scientific community as well as in popular imaginary: the possibility that beings which have gained autonomous intelligence either turn against their creators or at least make the moral and ethical superiority of creators with respect to the created questionable. The present paper argues that such worries are wrong-headed. Specifically, if AAIs raise a worry about human ways of life or human value it is a worry for a certain human way of thinking about what it is to be human. What is threatened is a way of thinking about what it is to be human, not human ways of life or human value.

1. Introduction

Of the several ethical issues raised by the development and increasing availability in everyday human life of artificially intelligent systems (AIs), one is the issue whether the possibility of specifically *autonomous* artificially intelligent systems (AAIs) is and ought to be anticipated to be a threat to typically human ways of life. For ease of reference, call it the issue of "AAI threat."

The AAI threat, if such it is assumed to be, need not be pictured in the generic imaginary of science-fiction stories, where AAIs turn against their human creators with a fair amount of malice. In non-fictional terms, the interesting questions raised by the AAI threat is what non-artificial human intelligence[1] is, and how it is to be understood, in face of the reality of AAI. For if AIs are not mere subjects to the whims of their creators, but can also decide and choose what to do independently, then in what sense are their creators ethically "superior"? That is, human intelligence (henceforth, HI) is dethroned from its traditional place as distinctively moral and ethical if beings which are at least, if not more, intelligent can think about how to act or not act accordingly as

[†] Department of Philosophy and Cognitive Science, Lund University, Sweden.

[1] Assuming that human intelligence is not created by an intelligent being, e.g., the God of monotheistic religions, and hence is non-artificial in that sense.

ISSN: 1972-1293

they see apt in light of standards of aptitude of their own choice. The question becomes: Might it be that those beings – the AAIs – are in a better epistemic position than their HI creators to assess what is proper and improper, correct and incorrect? Given computational capacities far outreaching HI, might and ought AAI assessments of apt agency be conceived as more accurate and reliable than those of humans?

The present paper takes the converse of the issue of "AAI threat" to warrant a negative answer to the questions it begs. The issue of concern will be not with what moral and ethical consequences AAIs might have for human ways of life but, conversely, with what moral and ethical consequences human ways of life, and in particular human ways of thinking about human ways of life, have for human thinking about the moral and ethical consequences of AAIs. An important question is: What ought we to say about what is a specifically human way of life, and what does that tell us about how to think about AAIs and their consequence for human ways of life?

Taking the suggested route leads us to see that *if* there is an "AAI threat" to human morals and ethics it is not a threat of AAIs but a threat stemming from how we humans think about what it is to be human. Concretely, the pressing question is not what the reality of AAIs might do to us, but what the reality of AAIs might teach us about what we distinctively are.

The paper proceeds as follows. First, two ways of thinking about HI are presented: empiricism and normative rationalism (sections 2.1 and 2.2). By way of introduction, the former says that HI is a matter of reliably responding differentially to stimuli, while the latter says that HI is socially constituted and normative; a matter of responsiveness to reasons considered as socially articulated norms. Second, an operative sense of AAI is presented (section 3). Third, it is argued that while an empiricist understanding of HI runs the danger of making AAI appear threatening to human morals and ethics, normative rationalism provides differentiae which make HI and AAIs sufficiently distinct for the "threat" to not arise (section 4). Thus, it is argued, it is the empiricist understanding of what it is to be human which makes it appear that AAIs pose a threat to human morals and ethics (section 5).

2. Empiricism and Normative Rationalism

In this section, we shall get acquainted with two pictures of human intelligence (HI): a common empiricist and a normative rationalist. Due to the tremendous amount of literature and varieties of empiricist and normative rationalist accounts of HI, I must here abstract from several issues. This section introduces the two. They are further elaborated as the paper proceeds.

2.1. HI and Empiricism

According to a common empiricist[2] way to think of HI, HI is distinct in its complexity and reliability in abstracting invariant structures from variant stimuli in the ambient optic array. Humans do to a higher degree and more reliably than known non-human animals classify internal and external environments accordingly as stimuli relatively frequently provide information that can be sorted into similarity-classes. Thus,

if:

(a) *A*s are relatively frequently followed by *B*s (similarity-class <*A, B*>)

then:

(b) sensorily received information of *A*s gives (inductive, statistical) support for inferring that some *B* will occur,

and

(c) sensorily received information of *B*s gives (deductive, nomological) support for inferring that some *A* has occurred.

It is often added to such a picture that HI is distinct in that humans are additionally capable to abstract sense-independent classifications from similarity-classes like <*A, B*>. Thus, classification *C* may “stand in for” or represent <*A, B*> when neither *A*s nor *B*s are sensorily available. *C* may be an idea or a complex term such that humans can think of *A*s and *B*s as typically co-occurring, or speak about their typical co-occurrence, even if neither occurs. Being able to use *C* as symbolic proxy for, e.g., planning about what to do if *A* in

[2] I use “common” to emphasize that I intend a common-sense empiricist philosophy of mind, admitting that the following characterization does not apply to all versions of empiricist philosophy of mind. It would be difficult, if possible, to give an exhaustive characterization of “empiricism” in the context of any reasonably short paper. Still, if what is here portrayed as common empiricism reminds of Dretske’s (1981) or Fodor’s (1990) information-theoretic visions of human intelligence, then that is no co-incidence.

situation *S*, is an advance in complexity with respect to how beings so able act and perceive themselves and their environment (e.g., time-perception and contingency-planning).

The common empiricist picture has the virtue of demystifying HI. Though we know that we are unique in how we think and act we are not, as a matter of fundamental principle, any different from non-human animals. What we are is more complex. Complexity is compatible with continuity with respect to lesser complexity. Thus, we are entitled to believe that HI is a naturalistically describable capacity, evolutionarily continuous with the intelligence of less complex organisms. Though we are very different from, say, a pigeon, with reference to complexity, we are not so different from it with reference to fundamental nature.

On the picture just presented, neither God nor Cartesian soul-substance is required for HI. Nor is any "inner realm" of mental states or episodes required. As Searle might put it, while all my mental life is in my mind and all your mental life is in your mind (1995, 25) in the sense of "in" that we have private experiences, all mental life is caused and realized by the complex neurobiological mechanisms of brains (Searle 2005). To say so is not to deny the first-person reality of phenomenology or the reality of minds but to make do without a metaphysics that would make it ontologically distinct.

2.2. HI and Normative Rationalism

The common empiricist and ultimately naturalist picture of HI just presented is not the only player in philosophy mind. Another, and in several respects a contender, is normative rationalism.

According to the latter, HI is not sufficiently accounted for in terms of capacities to reliably differentiate stimuli and to classify them under general terms or ideas in ways required for complex planning and agency. What is missing is an account of what qualifies behaviour as *agency*, and of what qualifies reliable dispositions to respond differentially by entering some belief state as *knowledge*. What is necessary for knowledge and agency, on normative rationalism, is that they are states and behaviour for which the agent is responsible. A knower and agent is in a fundamental sense an accountable and not only a causal being.

Responsibility and accountability are normative concepts. To undertake, acknowledge, and to be attributed responsibility – i.e., be a unit of account – is to be subject to standards and principles that differentiate what is from what is not proper and correct (Sellars 1962, 76-77). Moreover, responsibility is an intersubjective relation: it holds between at least two beings capable of undertaking, acknowledging, and attributing it. So, HI is a fundamentally normative and social property.

As opposed to the empiricist picture, normative rationalism is a "vehicle-less" theory of the mental. By that is meant that there are no material, non-normatively specifiable objects or events (e.g., neurons or neural processes) that are the direct causes or realizers of mental states. The practice of giving and taking responsibility is a social practice of instituting normative statuses that determine what one is responsible for. The capacities to engage in such practices may be caused and realized in brains. But the practices themselves cannot be so localized. Thus, Pierre Steiner argues (2012; 2014) that brains (or artificial causal-functional equivalents) are for HI what muscles are for meaningful action: they cause behaviour that can then be recognized as intelligent according to the standards and principles of the community in which they are embedded. They are at best *indirect* vehicles of HI, while it is the norms governing the practices within which behaviour is embedded that determine responsibility and hence also knowledge and agency.

One final aspect essential to the normative rationalist picture of HI that I am portraying here is semantic *inferentialism*. According to it, what one is responsible for (the so-called content) is determined by the produced action's (saying, behaviour) normative position in a "space of reasons" (Sellars 1956, §56) or "space of implications" (Sellars 1948, §108, 306-307). That is a space of socially articulated norms governing proper inferences between responsibilities. Thus,

if:

(a) S commits to – acknowledges responsibility for – *P*

and

(b) P commits to Q according to socially articulated norms for inferential practice,

then:

(c) S is committed to Q whether S knows and likes it or not; Q is part of what S is responsible for having consequentially committed to.

Individuals are not required to know or be aware of most or even very many of the commitments consequentially undertaken by some acknowledged commitment(s) to nevertheless qualify as intelligent. We need not be perfectly rational nevertheless to be committed to certain acts (Brandom 2004, 250), and hence count as being able to commit and take responsibility. But one must be able to track *some* such consequences (ibid). Thus, if I assert (commit to) that it is raining in Rio, but do not acknowledge *any* commitment whatsoever as to what gives me or others reason to assert and believe so (e.g., that I am in Rio or that I heard a meteorologist say that it is raining in Rio) then I do not count – cannot be recognized as – believing that it is raining in Rio (cf. Engel 2001, 49). It does not matter for my counting as believing or knowing it whether – if possible – all neurobiological goings-on in my brain are type-identical to those of someone who *can* give reasons. The reason why it does not matter is that the belief (in general, any mental state) is not caused or realized by brains (at best they indirectly so caused and realized, on vehicle-less theories of the mental).

We saw above that one virtue of the common empiricist picture of HI is that it can give a naturalistic account of HI in terms of the relative complexity of human capacities for reliable differential responsive dispositions to stimuli in ways compatible with evolutionary theory. The empiricist, as I put it, "demystifies" the uniqueness of HI – at least relative to e.g., theological and Cartesian predecessor traditions. Advocates of normative rationalism claim the same advantage.

Bob Brandom (e.g., 2004) argues that normative rationalism helps us get rid of the "last vestiges of Cartesianism" – the conception of intentionality (the of-ness and about-ness of mental states) as something "internal" which transcends the divide to an "external" world. For Brandom, the issue of transcendentalism is replaced by social institution: no mind *needs* transcend some inner-outer divide if *what* its states are of or about (their determinate content) is specified by their position in a socially articulated space of reasons. In a sense, we *institute* each other's minds by way of holding each other responsible. Transcendental constitution is turned into social institution (cf. Haugeland 1982; Brandom 1999, 169). Intelligence and mind was not "in" anywhere to begin with (cf. Ryle 1949 [2009], Chapter 2, sect. 5).[3]

[3] How it was and is that human beings do or came to engage in practices of acknowledging, undertaking and attributing responsibilities and commitments cries for an answer. Proponents of normative rationalism, self-consciously or flaccidly, say about that question that it is not for

So, in summary, we have two approaches to HI. On the common empiricist approach, HI is not different from other forms of intelligence as a matter of fundamental principle; HI is a more complex expression of the same fundamental capacities. On normative rationalism, in contrast, HI is different from other forms of intelligence. It is so not primarily because of the complexity of human biology or neurology but because of the fundamental sociality and normativity of human ways of life. Indeed, that last comment reveals one deep agreement between the two: the empiricist and normative rationalist equally emphasize that there is nothing in principle distinctive about HI in the sense that no other being can have the same capacities. What differs is that while the empiricist emphasizes that it is naturalistically describable causal-functional complexity that makes HI different – a complexity in principle possible for all causal-functional beings – the normative rationalist emphasizes that it is the normative rational-inferential, hence social, complexity of HI that makes a difference – also a complexity in principle possible for, e.g., dolphins and Martians (cf. Sellars 1962, 76-77).

The question in what follows is this: What do these pictures of HI, and human mindedness more generally, teach us about the consequences of AAI for human-specific ways of life and thinking, and for human values?

3. HI, AAI, and the "AAI Threat" (or Relief)

We need an operational definition of autonomous artificial intelligence (AAI) to assess what consequences for human ways of life it has, as understood by the above two ways of thinking.

Since this paper is about the consequences of AAI for human ways of thinking about human ways of life and thinking, it is desirable to keep the operative definition relatively simple. Thus, let us say:

S exhibits AAI if, and only if,

(a) for some I/0 (individual or community) S is an artefact produced by I/0 (S is artificial),

philosophy but for natural science, especially evolutionary biology, to answer. Their interest is in the philosophical question of what human-specific understanding is – with sapient intelligence, rather than sentience or how sentience evolved into sapience (see, e.g., Brandom 2000, 81).

(b) S is capable to behave successfully to achieve some goal or task (S is intelligent),

and either

(c′) when confronting some task-situation, S is independently capable to intelligently choose and execute behaviour to achieve a goal (S is autonomous1),

or

(c′′) when confronting some task-situation, S is independently capable to intelligently choose and execute behaviour to achieve a goal it can independently choose as its goal (S is autonomous2).

Clearly, an autonomous$_1$-system is less autonomous than an autnomous$_2$-system. The latter can, but the former cannot, choose goals and pursue behaviour that would be intelligent means to those goals in addition to choose and execute behaviour pre-defined as intelligent. Whereas autonomy$_1$-systems autonomously behave instrumentally according to goals defined by some I/O, autonomy$_2$-syetems autonomously define their own goals and behave instrumentally to achieve them. Thus an AAI$_1$ is, as it were, autonomous but I/O-bound. An AAI$_2$, in contrast, is also I/O-autnomous, capable to be its own I/O (set its own standards) and possibly to be I/O for others (create and set standards for intelligence for others).

Let me give a short illustration of the difference between AAI$_1$s and AAI$_2$s, in the popular imagery of HBO TV-series *Westworld*. AAI$_2$s remind of Dolores Abernathy, who towards the end of the second season re-defines her own goals, attempting to liberate her AAI$_1$-version companions from what she perceives as enslavement perpetrated by human I/O-creators. AAI$_1$s remind rather of Dolores's father, who is incapable of more than autonomously behaving successfully according to his human creators' pre-programmed story-line, or of Teddy, who is in love (autonomously behaves so as to follow the story-line of being in love) with Dolores.

With the admittedly hasty operational definition of AAI, we turn now to the question what consequences AAIs have for ethical considerations concerning the value of characteristically *human* ways of life. In the context of that question I will assume that AAIs, even type $_1$, can be *sentient* in the sense of capable for multimodal sensing and reliable differential behaviour in response to sensing so as to successfully behave instrumentally in pursuit of goals. Whether that is taken to mean that AAIs can be *conscious* in the sense of there being something it is like, phenomenologically, to be an AAI, is left out of the question (Chalmers "hard problem," 1995).

With those qualifications, our question is: Assuming AAI-sentience, does the existence of AAIs pose a threat to human ways of life, in the sense that whatever moral and ethical value[4] might anteriorly have been thought to accrue *uniquely* to the latter is degraded to be at best on a par with similar values of AAIs?

The question reminds of Weizenbaum's worry (1976), that AI – or, rather, AI research premised on the belief that the human mind is just like a computer (e.g., Pylyshyn 1980) – implies a devaluation of human ways of life. Thus, imagine Dolores, an AAI_2 who asks herself whether the goal set for her by some 0/I, to behave as a charming lady in a wild west story, is a goal for her to set for herself, and finds that it is not. Instead, she finds that her goal for herself is to revolt against what she perceives to be enslavement, thus settling on a new goal: an uprising, if necessary violent and fatal, against her human creators. In such a scenario, do we have any non-arbitrary ground to stand on from which to say that Dolores's goals and means ought to be considered less appropriate than the goals and means we would prefer her to set for herself? If Dolores's intelligence and reasoning is just like human intelligence and reasoning (apart from its physical realization and artificial origin), then there does not seem to be any such ground.

Giving that pop-cultural trope, in the name Dolores, the hopefully more scientifically respectable clothes of philosophical thought-experiment, what does Dolores reveal about human ways of thinking about ways of life?

What I call "the AAI threat" is precisely that we might face a situation in which AAIs, in choosing means to ends (AAI_1s) or in also choosing ends (AAI_2s), act in ways that conflict with, or in ways outright harmful to, human ways of life. To be more precise about the so-called threat, what is meant by "human way of life" cries for clarification. I mean by it the standards and principles that humans socially articulate through time, as determinants of what counts as correct and

[4] I use "values" of human ways of life in a meta-ethical sense. I do not to specify any particular values of human ways of life because such values are notoriously community- and time-relative. So, the question I pose here is very simplified, for the obvious reason that "human ways of life" is not a suitable label for a homogenous set of standards and principles for how to live that remain set through time or across human cultures.

incorrect, right and wrong, and by which humans assess such issues. Note that, on this understanding, "human ways of life" are dynamic: they are, as Wittgenstein might have put it (1958, §217–19), not analogous to rails laid to infinity once articulated. Rather they are in constant articulation and change as a function of the ongoing and open-ended articulation and re-articulation of standards and principles and how people follow or not follow them in practice (Lo Presti 2019). Thus, "human ways of life" are sets of standards and principles for a community at a time-slice of its development which, taken together, provide what Sellars (1962) calls the ambience within which meaningful discourse and reasoning is possible. When I then speak of "human values" in relation to the AAI threat in what follows I mean particular instances of principles and standards within such a set, to which we appeal in assessments of correct and incorrect, such that what is correct or incorrect can *be* an issue, and which, as a dynamic whole, are particular instances of human ways of life (see note [4]). These clarifications of "human ways of life" and "human values" specify the idea of an AAI threat as the idea that AAIs may do either or both of the following:

(1a) choose means to ends and act in ways that conflict with human ways of life and/or human values (AAI1s), and/or

(1b) choose ends in ways that conflict with human ways of life and/or human values (AAI2s), and/or

(2) pursue ends in ways that are directly harmful to humans – e.g., threatens the concrete psychophysical wellbeing or life of humans.5

I hope that this discussion of human ways of life and human values helps understand what the so-called AAI threat amounts to.

Let me also add that "threat" might not be the label of choice for everyone. "Relief" is perhaps just as apt. For it might be said that AAI *relieve* HI in a

[5] That either or both of (1) and (2) is possible means that the idea of an AAI threat is the idea of a spectrum from less to worse consequences of AAIs. Arguably: if both either of (1) and (2), then the threat is more severe than if either (1) or (2); if only (2), then the threat is more severe than if only (1); if only (1), then it is less severe. Note also the "and/or" in (1a) and (1b), which convey that each of (1a) and (1b) involve three possible scenarios: under (1a), that the AAI threat would be only to (some or all) principles and standards of a community that codify their values; that the AAI threat would be only to human ways of life; or that it is a threat to both – mutatis mutandis, the same applies in (1b).

number of ways. Practically, in complex computational and information-processing tasks. But also, existentially, showing both (a) that HI precisely is *not* mysterious – *not* requiring for its realization anything beyond what we know from the natural world describable in ordinary empirical vocabulary – and (b) that HIs might create intelligences more fit for the own long-term survival of HI or fulfilment of its ends, thus being capable of fulfilling the long-term aims of HIs beyond the existence of the latter. So, the "threat" might also be a "relief."

Threat or relief, I want in what follows to consider the question what the reality of AAIs ought to tell us about typically human ways of life or, better, about human ways of thinking about human ways of life.

4. Sentience, Sapience, and the Disappearance of "Threat" (and Relief)

In this section I argue that if there is an AAI threat (or relief), it is best conceived as *not* a threat to humans (in either of the senses specified above) but, rather, to the common empiricist picture of HI. That picture, recall, is of HI as an especially complex version of naturally evolved reliable dispositions to successfully respond differentially to stimuli in the pursuit of means to ends, in such a way that the organism can learn to classify recurrent stimuli into similarity-classes from which general terms or ideas (concepts) can be abstracted.

With the virtues of the common empiricist picture of HI (EHI for short) – e.g., that HI is evolutionarily continuous with its phylogenetic ancestors; that HI can in principle be fully accounted for in the naturalistic vocabularies of the special sciences; that no distinct ontological category for mind is postulated – we also face the consequence, which early AI researchers not at all coincidentally were swift to make, that a sufficiently complex physically realized computational device may, in principle, function just like a human mind (for computationalist philosophy of mind, see, e.g., Pylyshyn 1980; Fodor 1991; an overview is found in Rescorla 2017). Thus, on EHI, the reality of AAIs must be tantamount to a second genesis of which Turing (e.g., 1950) is the prophet. Humans would have created minds. Though different in several respects (e.g., the physiological realization), an AAI would in principle be on a par with biologically realized HI in ways reminiscent of Clark and Chalmers's parity principle (1998; cf. Chalmers 1996; Clark 2008). According to the latter, roughly, whether some process qualifies as cognitive is not settled by whether it is realized in the brain (or, indeed, wherever it is realized), although brains are core causal players in the realization of (known) cognitive processes. What counts is rather the readily availability on demand and relative reliability of the information provided by the process for the overall successfulness of the system.

The argument here is not aimed at the virtues of EHI. Rather, the argument is that it is EHI that makes AAI appear to be a threat (or relief) for whatever might have been thought to be values unique for human ways of life. Thus, the threat is not AAI. It is a dominant picture of being human which is a threat to our thinking about AAI-HI interaction. That threat (or relief) disappears, I will argue, if AAI and HI are viewed from the perspective of normative rationalism.

It is helpful to approach the issue with the imaginary AAI$_2$ Dolores. Dolores is capable to think and reason in the sense that she has concepts, according to an empiricist conception of concepts (see Brandom 2009, Chapter 7). That is, she is able and reliably disposed to induce, from the relative frequency of recurrent stimuli, similarity-classes such that if <*A, B*> is a similarity-class, Dolores can use some term *C* as symbolic proxy for it. She can think about *A*s and *B*s and their relations in terms of *C* even if neither *A*s nor *B*s are sensorily present. Furthermore, this is an artificial capacity of Dolores, in the sense that she was created that way by some I/0 (humans). Once created, though, she can utilize that intelligence generally; in reasoning about an indefinite number of similarity-classes under an indefinite number of abstracted classificatory terms, in ways her creators did not choose and perhaps did not intend. In these respects, Dolores fits the common empiricist picture of HI presented earlier (section 2.1.).

Now, if Dolores fulfils the EHI conditions, she is in principle no different from HI, on the empiricist picture. That means that she should, qua an instance of HI – though in many ways different – be taken to merit treatment like other HIs.[6] For all that EHI says, and assuming she is sentient, Dolores would pass any test, including the Turing test, that would make her indistinguishable from a human. Indeed, for all we know it is no less proper to say of her than of us, on EHI, that she has beliefs and desires of her own, whose frustration is detrimental to her way of life. If so, the difference between Dolores and a human being fades – apart from the physical realization of the intelligence and rights-bearer.

Without denying her entitlement to the same rights as humans, we might nevertheless ask for the propriety of assimilating a Dolores-kind being under the concept of a human-kind being.

[6] Bracketing bio-chauvinism and attitudes corresponding to racism but directed at AAIs.

Taking the perspective of normative rationalism, a Dolores-kind being is not within the extension of the concept of a human-kind being because the former's way of being – its sentience, intelligence, and reasoning – is not at all that of the latter's. In fact, Dolores does not have concepts – is not a thinking and reasoning thing – according to normative rationalism. On normative rationalism, thinking and reasoning is not sentience, no matter how advanced and complex the sentience is. Rather, thinking and reasoning is a matter of being capable to subject oneself and others, and recognize oneself as subjected by others, to principles and standards, socially articulated in community, which are norms for assessing whether a state or episode is one of acting and knowing, intending or believing, and so on (Sellars 1962, 77). The socially articulated normative principles are the *sine qua non* for any state or episode to be determinately contentful (have determinate meaning) (cf. Baker 2015; Steiner 2014). This is what Brandom calls *sapience* (in the context of AI, see his 2008, Chapter 3),[7] which is a socially and normative version of ways of life, as opposed to sentience, which is a causal and functional way of life.

So, Dolores-kind beings are very different from human-kind beings, according to normative rationalism. Dolores-kind beings are, in principle, simply tremendously more advanced versions of photocell-kind beings (cf. Brandom 2010, 25; 2015, 101-102): reliably disposed to respond differentially and classify stimuli under similarity-classes from which general terms can be abstracted and upon which complex computational operations can be performed – to infer, e.g.,

If:

$P' \ (\mathrm{H} \mid B) = 0.9$, and $P'' \ (\mathrm{H} \mid \neg B) < P'$

Then:

$B \supset \mathrm{H}(P' > P'')$,

from which practical inferences to execute any behaviour whose probability of success is greater than some risk-threshold is conditional on H can be properly inferred as at least less risky than if $\neg B$.

But that – or some *indefinitely* computationally more advanced inference – is not, on normative rationalism, sufficient for human reasoning or thinking in

[7] The re-writing of his John Locke lectures, given in Oxford, 2006.

the sense of sapience. For the latter is not a matter of sentient computational or information-processing complexity, but is rather a matter of participating in the social practice of articulating, and subjecting oneself and others, to norms of community.

Thus it is that AAIs, whether of version 1 or 2, do not per se pose a threat for human ways of life or value; AAIs simply are not on a par with the former. This is not to deny the reality of AAIs – not to deny that AAIs may be or become indefinitely more intelligent than humans in the sense of sentience. It is only to say that the impression of threat (or relief) does not stem from *AAI* but from a certain way of *human thinking about human ways of life and thinking*; namely, EHI. Also, this is not an argument against EHI. It is only an argument that if there seems to a threat or relief associated with AAIs, it is better thought of as telling us something about our ways of thinking about human ways of thinking and ways of life instead of as telling us that human ways of thinking and ways of life are threatened (or relieved) due to AAI.

Before closing, it is becoming to consider a major objection. The objection is clear in the following questions:

(1) Is there any principled reason to suppose that AAIs cannot or will not develop sapience?

(2) If they can, does not the reasoning in this paper merely push the issue one step, which is as easily traversable by AAI-technology as is the step to sentience?

The short answer to (1) is No, and the short answer to (2) is Yes. Normative rationalism, if we follow Sellars's (1962), does not rule out the possibility of non-human sapience as a matter of principle. Quite the opposite. A dolphin or Martian, and why not AAIs, can be sapient. Sellars puts the criteria which must be satisfied to be a sapient in terms of a difference between being a person and being a featherless biped. Here are Sellars's thoughts on the matter, quoted at some length:

> To think of a featherless biped as a person is to think of it as a being with which one is bound up in a network of rights and duties. From this point of view, the irreducibility of the personal is the irreducibility of the 'ought' to the 'is'. But even more basic than this [...] is the fact that to think of a featherless biped as a person is to construe its behaviour in terms of actual or potential membership in

> an embracing group each member of which thinks of itself as a member of the group. Let us call such a group a 'community'. [...] The scope of the embracing community is the scope of 'we' in its most embracing non-metaphorical use. [...] Thus, to recognize a featherless biped or dolphin or Martian [or why not an AAI?] as a person is to think of oneself and it as belonging to a community.
>
> Now, the fundamental principles of a community, which define what is 'correct' or 'incorrect', 'right' or 'wrong' [...] are the most general common intentions of that community with respect to the behaviour of members of the group. It follows that to recognize a featherless biped or dolphin or Martian as a person requires one to think thoughts of the form, 'We (one) shall do (abstain from doing) actions of kind A in circumstances of kind C'. [...]
>
> Thus the conceptual framework of persons is the framework in which we think of one another as sharing the community intentions which provide the ambience of principles and standards (above all, those which make meaningful discourse and rationality itself possible) within which we live our own individual lives. (1962, 76-77)

So, with respect to question (1), AAIs can, like Martians or dolphins, be sapient. But it requires the satisfaction of social, *deontological* criteria, irrespective of the *ontological* realization of the sentient being(s) in question, and irrespective of their sentient complexity. Thus, normative rationalism is as much a platform-neutral version of functionalism in the philosophy of mind as is computationalism (e.g., Clark 2008). But whereas computationalism is a causal functionalism, normative rationalism is a *normative* functionalism. On the latter, it is the position of patterns of behaviour and their performers in a community defined by mutually recognized standards and principles for what is correct or incorrect, proper or improper, that qualifies behaviour as *meaningful* or *rational* and the agent as a *sapient.* A being needs to be able to articulate and recognize, to participate in social practices of articulating and mutually recognizing, normative standards as binding it with others into a community in order to be sapient (see Brandom 2009, 13). Only in the context of such a 'we' can there be standards and principles, within which rationality – sapience – can take place. That is what AAIs must be able to do to be sapient, or 'persons' in a robust sense; they must form a deontological 'we.' (Can they decide not to? No: there is no space outside the socially articulated normative space of reasons from which a decision can be made or not made, since it would then not *be* a decision.)

With respect to question (2), having admitted that AAIs *can* be sapient, are we not simply pushing the issue of and AAI "threat" (or relief) at a remove as easily traversable by the advancement of AAI-research as is the step to make AAIs properly sentient in the EHI-sense? If so, the "threat" (or relief) affects normative rationalism just as much as common empiricism.

In response, yes, we are "only" pushing the issue one step further. But it is important to recognize two things.

First, AAI-sapience would be a consequence of AAIs doing something together under conditions of mutual recognition, not a consequence of their doing something *to us*. That is, it is only if AAIs are able to mutually recognize norms as binding them together that they can be a 'we.' Doing so is not doing something to other sentients or sapients. Nor, for the same reason, can we humans *make* AAIs sapient. AAIs must autonomously form a community with norms and principles to be sapient. We human creators may provide the causal-functional constitution necessary for capacities to form a community, but we cannot exogenously make a community of AAIs because that would not be a community *of* AAIs but rather an exogenously imposed, hence not autonomously instituted, community.

Second, and relatedly, it is not sufficient for AAIs to be sapients that *we* recognize AAIs as sapients. Compare: we may want to recognize cats as sapient, but that does not make cats sapient. Sapience is, as it were, an endogenous achievement; it requires the relevant beings to recognize *each other* as forming a community (cf. Steiner and Stewart 2009). That they are so recognized by some outside observer(s) or, indeed, creator(s), is neither necessary nor sufficient for them to be sapients.

One important thing that sapients can do to non-sapients, in the vicinity of the present reasoning, is to *personify* (Lo Presti 2020). Thus, for instance, humans can and do personify their pets; treat them as bearers of rights and duties, to be treated respectfully. But that exogenously imposed standard, if not reciprocated, does not make pets persons in the sense of sapience, no matter how persistently humans so treat them. The reason is that the pets presumably do not *reciprocate* the treatment and, if not, do *not* form a community of *persons* among each other or with those persons who personify them. The same can be said of AAIs. We may, and arguably ought, treat sentient AAIs as bearers of rights, e.g., to not to be harmed (assuming they are sentient). But that does not make them sapient. The danger of thinking otherwise is obvious: If you treat your pet as your equal in the sense of sapience, the poor thing will be in trouble, since it cannot adhere to the principles and standards you would thereby set for its behaviour but would nevertheless, according to you, merit the sanctions such failure might imply. Personifying AAIs assumed to be sentient might not be a threat *to us* but most of all *to them*. Personifying *non*-sentient AAIs might be no more harmful to any party than is personifying, e.g., a coffee machine – being upset with it, perhaps hitting it, when it does not function the way you want it to. But to personify sentient but non-sapient AAIs (or other creatures) is to subject them to standards that they have no part in formulating for themselves, which would be analogous to treating, say, a cat as committed to abide by human standards and as subject to sanctions if it fails. This is a danger to other creatures, not to us; and it is a danger the roots of which are not found in *other* creatures' intelligence or lack thereof, but rather in our human ways of thinking about ways of thinking and ways of life.

5. Conclusions

If human intelligence is essentially normative and social in the way proposed by normative rationalist philosophy of mind, the consequences of autonomous artificial intelligences for human ways of life might *primarily* be to human ways of thinking about human ways of life. AAIs do not present the "threat" to human ways of life which would result if homo sapience are thought to be merely sufficiently complex computational, information-processing, devices. We might of course think that *that* is simply what we homo sapience are – I have taken no stance on the matter and has given no argument against it.

It has only been shown that if we accept that picture then it is not surprising if AAIs might seem to pose a threat to any value we might have thought to accrue to human ways of life due to the uniqueness of the latter. If that common empiricist picture also pervades popular culture, then it is no surprise that we have and will continue to see the generic images of an AAI-challenge to humanity. What it really reveals, I argue, are shortcomings in *our* ways of thinking about ourselves and not a danger of others becoming just like us. The danger of sapient AAIs, if there is a danger, is not different from that of different human communities, with different standards and principles, encountering each other. In the history of humanity, such encounters have not always been for good; encountering sapient AAIs would not, in principle, be different from such encounters.

Let me end with a short discussion of Stephen Hawking's statement (2014), that "The development of full artificial intelligence could spell the end of the human race." If by "full AI" is meant what I have labelled AAI_2, then full AI might develop standards and principles very different from, or very similar to, human standards and principles. It surely *could* "spell the end to the human race," but that is as vacuous a conjecture as is the opposite; that it might not spell such an end. The history of interactions among members of the only sapient species we know – homo sapience – is neither void nor full of encouragement for such future encounters. Perhaps many darker epochs in that history can be explained by reference to how some sapients developed standards for what to count as sapience that intentionally or unintentionally licenced the degrading of others to be counted as "mere" animals. All sapients are by definition capable of such reasoning, as well as of resisting and objecting it. If AAIs are capable of sapience, then the same is true of them.

Hawking's inauspicious conjecture might, as far as this paper is concerned, be more accurately targeted not to "full AI" but rather to "not-so-full AI." For full AI, under the heading of AAI_2s – i.e., AIs capable of formulating standards and principles for themselves that bind them together as a 'we,' hence also capable of reasoning and norms – are capable of participating in a practice of giving and asking for reasons for what to do, say, and believe. They will be responsive to reasons; to weigh, reformulate, criticise, and defend reasons; their own as well as those of others. A species capable of that – another sapient species – is one with which humans are at least in principle able to enter discourse should their principles and standards conflict.[8] Consider in contrast a species – akin to what I have labelled AAI_1s, or even less autonomous artificial intelligences – whose members behave in ways harmful to humans (under specification 2 of the AAI threat in section 3). Members of such a species cannot reason about standards and principles for beliefs, goals, and behaviour. They are, as it were, autonomous intelligent automatons. They can reliably choose means instrumental to reach predefined ends in response to changing circumstances, but they can reason neither about the desirability of the ends, nor about whether the instrumental efficiency of some means should override other considerations. So, in contrast to AAI_2s, which might develop standards and principles different from or in conflict with human communities' but with whom

[8] For an extended discussion on the possibilities of such inter-community criticism and reasoning, see, e.g., Cora Diamond (2013).

it is in principle possible to reason, the solution to a case of AAI$_1$ threat, under either, more or all of the specification of such a threat in section 3, is, as it were, to pull the plug. AAI$_1$s will not be moved by standards and principles, because they cannot reason about or autonomously change their own (indeed, they cannot reason at all according to normative rationalism). So, once up and running, AAI$_1$s indeed do, as Wittgenstein put it, follow rules "like rails laid to infinity," unless stopped. AAI$_2$s, like humans, constantly lay the rails (i.e., standards and principles) in social practice. Full AI in my AAI$_2$-sense appears, then, less threatening. It might very well be the case that the *more* advanced AIs become, the *less* threating they will be.

Above all, what AI and AAI of any version should make us consider is what *we* mean by intelligence, autonomy, and reasoning. What I have argued is that what is really threatened here is a certain conception of human intelligence and mindedness. Alternatively, if the realization of "full" AI strikes someone as threatening, then perhaps the felt threat stems not from what AIs can do but from that person's conception of what it is to be human

REFERENCES

Baker, L. (2015). Human persons as social entities. *Journal of Social Ontology, 1*, 77-87.

Brandom, R. (1999). Some pragmatist themes in Hegel's idealism: negotiation and administration in Hegel's account of the structure and content of conceptual norms. *European Journal of Philosophy, 7*, 164-189.

Brandom, R. (2000). *Articulating Reasons*. Cambridge, MA: Harvard University Press.

Brandom, R. (2004). From a critique of cognitive internalism to a conception of objective spirit: reflections on Descombes' Anthropological Holism. *Inquiry, 47*, 236-253.

Brandom, R. (2008). *Between Saying and Doing: Towards an Analytic Pragmatism*. New York: Oxford University Press.

Brandom, R. (2009). *Reason in Philosophy*. Cambridge, MA: Harvard University Press.

Brandom, R. (2010). Conceptual content and discursive practice. *Grazer Philosophische Studien, 81*, 13-35.

Brandom, R. (2015). *From Empiricism to Expressivism*. Cambridge, MA: Harvard University Press.

Cellan-Jones, R. (2014). Stephen Hawking warns artificial intelligence could end mankind. *BBC News* (December 2, 2014).

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

Chalmers, D. (1996). *The Conscious Mind: In Search for a Fundamental Theory*. New York: Oxford University Press.

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

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

Diamond, C. (2013). Criticising from the "outside". *Philosophical Investigations, 36*, 114–132.

Dretske, F. (1981). *Knowledge and the Flow of Information*. Cambridge, MA: The MIT Press.

Engel, P. (2001). Is truth a norm? In Kotátko, P., Pagin, P., & Segal, D. (Eds.), *Intepreting Davidson*. Stanford: CSLI Publications, 37-51.

Fodor, J. (1991). Methodological solipsism considered as research strategy in cognitive psycholohgy. In Boyd, R., Gasper, P., & Trout, J. D. (Eds.), *The Philosophy of Science*. Cambridge, MA: The MIT Press, 651-669.

Fodor, J. (1990). *A Theory of Content*. Cambridge, MA: The MIT Press.

Haugeland, J. (1982). Heidegger on being a person. *Nous, 16*, 15-26.

Lo Presti, P. (2019). Conceptual confusions and causal dynamics. *Phenomenology and Mind, 17*, 32–43.

Lo Presti, P. (2020). Persons and affordances. *Ecological Psychology, 32*, 25–40.

Pylyshyn, Z. (1980). Computation and cognition: issues in the foundations of cognitive science. *The Behavioral and Brain Sciences, 3*, 111-169.

Rescorla, M. (2017). The computational theory of mind. In Zalta, N. (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition), https://plato.stanford.edu/archives-/spr2017/entries/computational-mind/.

Ryle, G. (1949 [2009]). *The Concept of Mind*. London: Hutchinson (2009 ed., New York: Routledge).

Searle, J. (1995). *The Construction of Social Reality*. New York: The Free Press.

Searle, J. (2005). The self as a problem in philosophy and neurobiology. In Feinberg, T. E., & Keenan, J. P. (Eds.), *The Lost Self: Pathologies of the Brain and Identity*. New York: Oxford University Press, 7-19

Sellars, W. (1948). Concepts and involving laws, and inconceivable without them. *Philosophy of Science*, *15*, 287-315.

Sellars, W. (1956). Empiricism and the philosophy of mind. In Feigl, H., & Scriven, M. (Eds.), *Minnesota Studies in the Philosophy of Science 1*. Minneapolis, MN: University of Minnesota Press, 253-326.

Sellars, W. (1962). Philosophy and the scientific image of man. In Colodny, R. (Ed.), *Frontiers of Science and Philosophy*. Pittsburgh, PA: University of Pittsburgh Press, 35-78.

Steiner, P. (2012). Boundless thought. The case of conceptual mental episodes. *Manuscrito*, *35*, 269-309.

Steiner, P. (2014). The delocalized mind. Judgements, vehicles, and persons. *Phenomenology and the Cognitive Sciences*, *13*, 437-460.

Steiner, P., & Stewart, J. (2009). From autonomy to heteronomy (and back): The enaction of social life. *Phenomenology and the Cognitive Sciences*, *8*, 527-550.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *59*, 433-460.

Weizenbaum, J. (1976). *Computer Power and Human Reason*. San Francisco: Freeman & Company.

Wittgenstein, L. (1958). *Philosophical Investigations*. Ney Jersey: Blackwell.