Received: 17 January 2021 | Revised: 17 May 2021 | Accepted: 20 June 2021

DOI: 10.1111/rati.12320

ORIGINAL ARTICLE

WILEY

Existential risk from AI and orthogonality: Can we have it both ways?

Vincent C. Müller[1,2,3] | Michael Cannon[1]

[1]Faculty of IE/IS, Philosophy and Ethics Group, Eindhoven University of Technology, Eindhoven, the Netherlands

[2]School of Philosophy, Religion and History of Science, Interdisciplinary Ethics Applied (IDEA) Centre, University of Leeds, Leeds, England

[3]The Alan Turing Institute, London, UK

Correspondence
Vincent C. Müller, Faculty of IE/IS, Philosophy and Ethics Group, Eindhoven University of Technology (TU/e), PO Box 513, De Zaale, Atlas 9.329, NL-5600 MB Eindhoven, the Netherlands.
Email: v.c.muller@tue.nl

Abstract

The standard argument to the conclusion that artificial intelligence (AI) constitutes an existential risk for the human species uses two premises: (1) AI may reach superintelligent levels, at which point we humans lose control (the 'singularity claim'); (2) Any level of intelligence can go along with any goal (the 'orthogonality thesis'). We find that the singularity claim requires a notion of 'general intelligence', while the orthogonality thesis requires a notion of 'instrumental intelligence'. If this interpretation is correct, they cannot be joined as premises and the argument for the existential risk of AI turns out invalid. If the interpretation is incorrect and both premises use the same notion of intelligence, then at least one of the premises is false and the orthogonality thesis remains itself orthogonal to the argument to existential risk from AI. In either case, the standard argument for existential risk from AI is not sound.—Having said that, there remains a risk of instrumental AI to cause very significant damage if designed or used badly, though this is not due to superintelligence or a singularity.

KEYWORDS

existential risk, general intelligence, instrumental intelligence, orthogonality, singularity, superintelligence

1 | EXPOSITION: FROM SUPERINTELLIGENT AI TO EXISTENTIAL RISK

1.1 | The singularity claim

This remarkable statement from 56 years ago summarises the idea of the singularity and its risk:

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion', and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. (Good, 1965, p. 33)

In this vein, authors like Bostrom (2014) and Russell (2019a) claim that artificial intelligence (AI) is on a trajectory of development that will reach up to systems that have a roughly human level of intelligence. It has been claimed that this level has a 50% probability of occurrence by 2040–50 (Müller & Bostrom, 2016). When this level is reached, these systems would have the ability to self-improve, or to develop further AI systems, and thus surpass the human level of intelligence, reaching 'superintelligence'. The point in time where superintelligent AI is reached is often called 'the singularity' (Kurzweil, 1999, 2005, p. 487). After the singularity, developments are largely out of human control because it is hard to control entities which are more intelligent than humans. We summarise these thoughts in a claim:

> *Singularity claim:* Superintelligent AI is a realistic prospect, and it would be out of human control.

Given the lack of human control, developments after the singularity may well take a course that is negative for *homo sapiens*, even leading to the extinction of our species; this is called an 'existential risk'. So, from the singularity claim some people conclude that AI poses an existential risk for humanity, while others, notably Kurzweil, have positive expectations for the future after the singularity. The discussion is summarised in (Chalmers, 2010) as well as (Armstrong, 2014; Eden et al., 2012; Müller, 2020; Shanahan, 2015; Yampolskiy, 2018). In our reconstruction of the argument to existential risk from AI, we identify two premises: The singularity claim, as just sketched, and the orthogonality thesis.

1.2 | Orthogonality—The thesis

It appears that the argument that superintelligent AI represents an existential risk also requires the premise that the superintelligent AI is not necessarily ethical or even super-ethical (Russell, 2019b, p. 51). If it were ethical, superintelligent AI would not constitute an existential risk—unless, of course, ending the existence of the human species *is* the most ethical course of action, in which case we should probably not call this a *risk*, however (we will return to this possibility below).

To support the conclusion of existential risk, the singularity claim is thus explicitly supplemented by a further premise:

> *The Orthogonality Thesis:* Intelligence and final goals are orthogonal axes along which possible agents can freely vary. In other words, more or less any level of intelligence could in principle be combined with more or less any final goal. (Bostrom, 2012, p. 73; 2014, p. 107).

To understand this terminology of orthogonality, we can imagine a Cartesian coordinate system, where the *x*-axis is the one dimension (say: intelligence) and the *y*-axis is the other (say: goals)—so the dimensions are 'orthogonal'. Each agent has a tuple of two values, *x* and *y*, and any combination of values is possible. These goals are often formulated in terms of allocating 'utility' to outcomes, also called a 'utility function'. As an illustration of orthogonality, one might think of various chess computers where the degree of intelligence is independent from the goal, which is the same for all systems. (While this is a useful metaphor, of course it is doubtful that goals or intelligence can be thought of as distributed in just one dimension.) We assume that the qualification 'more or less' in the quotation is not meant to imply ethical limits, but rather paradoxes like goals that imply reducing intelligence.

Bostrom explains: 'The orthogonality thesis implies that synthetic minds can have utterly non-anthropomorphic goals—goals as bizarre by our lights as sand-grain-counting or paperclip-maximizing.' (Bostrom, 2012, p. 753). So, intelligent agents can have a wide variety of goals, and any goal is as good as any other.

1.3 | Existential risk—The conclusion

What is the situation if we cannot guarantee that the superintelligent AI is 'docile', as per Good's formulation? It may well be that the AI pursues goals that lead to human extinction, either by design or as a side-effect (Bostrom, 2002, 2013). As Bostrom summarises in his popular book:

> Before the prospect of an intelligence explosion, we humans are like small children playing with a bomb. Such is the mismatch between the power of our plaything and the immaturity of our conduct. Superintelligence is a challenge for which we are not ready now. (Bostrom, 2014, p. 259)

A terminological note: In these discussions, there is talk of 'existential risk' for humanity, and there is also talk of 'catastrophic risk', which is a larger set that includes risks for events that would produce a very large and lasting damage, but perhaps not involving the end of the human species (Bostrom & Ćirković, 2011; Häggström, 2016; Ord, 2020; Rees, 2018). Catastrophic risks generally pose a particular challenge to our understanding because they combine very large damage with very low probability of occurring—while we are much more used to dealing with risks that have moderate damage and a moderate probability of occurring, such as a river flood (Ord et al., 2010). It follows from this observation that any discussion of the very high-impact risk of singularity has justification *even if* one thinks the probability of such singularity ever occurring is very small. Because superintelligent AI may have more or less any goal and is out of human control, there is a risk that the goals it pursues will put it at odds with humans.

From these rough indications in the literature, we reconstruct the basic argument to existential risk from AI as follows:

> Premise 1: Superintelligent AI is a realistic prospect, and it would be out of human control. *(Singularity claim)*
>
> Premise 2: Any level of intelligence can go with any goals. *(Orthogonality thesis)*
>
> Conclusion: Superintelligent AI poses an existential risk for humanity

2 | IS THERE A TRICK? TWO KINDS OF INTELLIGENCE

The reconstruction shows that the argument for existential risk from AI has two premises, namely the singularity claim and the orthogonality thesis. We do not wish to challenge either of these premises; in fact, we think they are true (under the right interpretation). We are also not just complaining that central terms are undefined, or that they are value-laden (cf. Cave, 2020). Our concern is with the *validity* of the argument: It appears that for each of the two premises to be charitably interpreted as true, they must be interpreted as using the term 'intelligence' in different ways. If that is the case, they cannot be combined as premises into a valid argument for existential risk from AI. If, on the other hand, the premises use the same notion of intelligence then, we will argue, one of the premises turns out to be false, and thus the argument is not sound.

In discussions of the first premise, the singularity claim, superintelligence is typically explained on the basis of *general human intelligence*, where 'super' intelligence is just *more of the same:*

> We can tentatively define a superintelligence as any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest … Note that the definition is noncommittal about how the superintelligence is implemented. It is also noncommittal regarding qualia; whether a superintelligence would have subjective conscious experience might matter greatly for some questions (in particular for some moral questions), but our primary focus here is on the causal antecedents and consequences of superintelligence, not on the metaphysics of mind. (Bostrom, 2014, p. 22)

In the following, Bostrom discusses a few types of superintelligence: Speed Superintelligence, Collective Superintelligence, and Quality Superintelligence (Bostrom, 2014, pp. 52–56). Furthermore, Bostrom adds an endnote: '… we make no assumption regarding whether a superintelligent machine could have "true intentionality" (pace Searle, it could; but this seems irrelevant to the concerns of this book).' (Bostrom, 2014, p. 265). So, the singularity claim assumes a notion of intelligence like the human one, just 'more' of it. For the time being, we can leave it to this vague characterisation we are given.

In the second premise, the orthogonality thesis, on the other hand, it appears that we are looking at *instrumental* intelligence, i.e., intelligence in the instrumental service of given goals:

> For our purposes, 'intelligence' will be roughly taken to correspond to the capacity for instrumental reasoning […]. Intelligent search for instrumentally optimal plans and policies can be performed in the service of any goal. (Bostrom, 2012, p. 73)

This fits the notion of intelligence that Legg and Hutter synthesised out of over 70 different definitions of intelligence: 'Intelligence measures an agent's ability to achieve goals in a wide range of environments.' (Legg & Hutter, 2007, p. 402). S. Russell says 'Roughly speaking, an entity is intelligent to the extent that what it does is likely to achieve what it wants, given what it has perceived.' (Russell, 2019a, p. 14). More precisely, instrumental intelligence stands in the tradition of classical decision theory, which assumes that a rational agent should always try to maximise expected utility—by evaluating the possible outcomes, allocating a (subjective) utility to these and estimating the probability of them occurring (e.g., Simon, 1955). The *expected utility* of each choice is equal to the *sum of the utility of the possible outcomes, multiplied by probability for each outcome*. Rational choice that involves *other agents* is modelled in 'game theory' for decisions with and without uncertainty since the 1940s (Neumann & Morgenstern, 1944). Note that this decision theory is normative, i.e., it says that rational humans *should* decide in this way, not that they actually *do* decide in this way—in fact we often do not, which means the theory has little use for predicting human behaviour.

This theory is traditionally assumed as a matter of course in AI:

> In short, *a rational agent acts so as to maximise expected utility*. It's hard to overstate the importance of this conclusion. In many ways, artificial intelligence has been mainly about working out the details of how to build rational machines. (Russell, 2019a, p. 23, also 44 & 93; cf. Russell & Norvig, 2020, ch. 12 & 16).

Another example, closer to our discussion, is the programmatic paper 'Superintelligence does not imply benevolence' by Fox and Shulman, where the general model for intelligent behaviour is the 'AIXI formalism, which combines Solomonoff induction with Bayesian decision theory to optimize for unknown reward functions' (2010, p. 1). In this kind of formalism, a reward function is set and then the system optimises for that function in a given environment. Higher intelligence is then a matter of optimisation power, achieving maximal reward function in a larger set of environments. As Fox and Shulman rightly say, AIXI and similar systems have no room for a revision of reward functions. In short: 'AI has adopted the standard model: we build optimising machines, we feed objectives into them, and off they go.' (Russell, 2019a, p. 172). It is crucial for orthogonality that the 'objectives', 'reward functions' or 'utility' are given to the system, and remain stable, no matter how good, or intelligent, the system becomes at reaching those objectives.

Indeed, the revision of utility functions seems to play a very minor role, even in normative decision theory. Vassend remarks 'it is … puzzling that even the very possibility of updating the utility function seems to have been completely neglected in the literature.', and continues, 'However, there is nothing in the formal theory that prevents one from instead updating the utility function, while keeping the probability function fixed.' (Vassend, 2021).

In the following we shall explain our observation that the current argument for the existential risk of superintelligent AI leverages two different notions of intelligence: In the singularity claim we imagine an agent 'like us', but more intelligent *(general intelligence)*, while in the orthogonality thesis, we imagine an agent that has an ability to find ways to reach a given goal, but who would not reflect on that goal *(instrumental intelligence)*.

The argument to existential risk from AI is thus to be understood as follows (with indices, g = general, i = instrumental):

> Premise 1*: Superintelligent^{g} AI is a realistic prospect and would be out of human control *(Singularity claim)*
>
> Premise 2*: Any level of intelligence^{i} can go with any goals *(Orthogonality thesis)*
>
> Conclusion*: Superintelligent^{g,i} AI poses an existential risk for humanity

We tried to find a charitable interpretation and think the premises are both true, *as interpreted here*. But the argument is now *invalid* since the premises use different notions of intelligence. The orthogonality thesis is itself orthogonal to the argument to existential risk. Note that this holds also if one takes the plausible line to interpret general intelligence as a superset of instrumental intelligence, because in this case the orthogonality thesis would only hold for a part of general intelligence, the instrumental part (but a superintelligent agent would have to have general intelligence for the singularity claim to hold). More on this in a moment.

Perhaps it helps to represent the tension we identify in a little table (Figure 1):

	Orthogonality	Existential risk
Instrumental intelligence	Consistent	Inconsistent?

	Orthogonality	Existential Risk
Instrumental Intelligence	Consistent	Inconsistent?
General Intelligence	Inconsistent?	Consistent

FIGURE 1 The tension

	Orthogonality	Existential risk
General intelligence	Inconsistent?	Consistent

The only way that this argument can be made to work is by making sure the *same* notion of intelligence is used in both premises. In the next section we explore the different combinations of the above to consider whether this can be done: Is there a notion of intelligence that we can use in both premises and with which we can interpret both premises as true?

3 | CAN WE HAVE IT BOTH WAYS?

3.1 | Orthogonality thesis & general intelligence?

In this section we discuss what happens if we combine orthogonality with *general* intelligence and thus interpret the orthogonality thesis thus:

Premise 2**: Any level of intelligence^{g} can go with any goals *(Orthogonality thesis)*

It appears that a general intelligence would be able to reflect on goals and possibly revise them in the light of rational thought. Humans, for example, can reflect on goals, and we can also reflect on our goals on *ethical* grounds and achieve ethical insight—so one would expect that an 'intellect that greatly exceeds the cognitive performance of humans' (Bostrom, 2014, p. 22) can do so, too. We often reflect on goals when a goal conflicts with others and we have to decide which goal is more important. Such reflection is often invoked in accounts of moral responsibility of human agents: It is standard to specify the conditions for human responsibility for an action in terms of two conditions, an *epistemic condition*, and a *control condition*. The epistemic condition specifies what the agent should have known at the time of action while the control condition demands that the agent, at the time of action, should be responsive to reasons for and against the action, including moral reasons (Fischer & Ravizza, 2000, pp. 28–91); this is sometimes formulated as saying that the action had 'The right kind of cause' (Sartorio, 2016, p. 109). The orthogonality thesis is thus much stronger than the denial of a (presumed) Kantian thesis that more intelligent beings would automatically be more ethical, or that an omniscient agent would maximise expected utility on anything, including selecting the best goals: It denies any relation between intelligence and the ability to reflect on goals.

Human non-theistic ethics often starts from observations of descriptive facts, e.g., that humans avoid pain (Bentham), aim for the good (Aristotle), or expect other humans to follow universal rules (Kant) and moves from there to a superior goal or ethical rule. Whatever one may think about the rationality of these attempts to establish ethical obligations, if a human had been brought up to have 'goals as bizarre … as sand-grain-counting or paperclip-maximizing', they *could* reflect on them and revise them in the light of such reflection. Humans are capable of imagining moral progress for themselves and for societies; they even seem quite capable of contemplating deeply transformative changes to a different set of goals, even though this poses epistemic challenges (e.g., on the life of a vampire, see Paul, 2014). Indeed, many humans show a constant reflection on ethics.

So, what would prevent a generally superintelligent agent from reflecting on their goals, or from developing an ethics? One might argue that intelligent agents, human or AI, are actually *unable* to reflect on goals. Or that intelligent agents *are able to* reflect on goals, but *would* not do so. Or that they would never *revise* goals upon reflection. Or that they would reflect on and revise goals but still not *act* on them. All of these suggestions run against the empirical fact that humans do sometimes reflect on goals, revise goals, and act accordingly. If one starts from a notion of *general* intelligence, orthogonality will need very substantial new arguments to become credible.

Since this road to a sound argument does not look promising, let us try the other option to use a single notion of intelligence in both premises: instrumental intelligence for both.

3.2 | Singularity claim & instrumental intelligence?

In this step, we assume that we have an agent that has only instrumental ability to find efficient ways to achieve goals, and we consider whether such instrumental intelligence is sufficient to become the kind of superintelligence which poses an existential risk—so here we investigate the consistency of the singularity claim as per 'Premise 1', interpreting the intelligence in question to 'instrumental intelligence':

> Premise 1**: Superintelligent[i] AI is a realistic prospect and would be out of human control *(Singularity claim)*

We shall grant that this kind of superintelligent instrumental agent *is* a possibility (as does Chalmers, 2010, fn. 20), despite the significant concern that a superintelligent instrumental ability would seem to require an understanding of the world that includes understanding agents, intentions, and ethical reflections on goals. This thought it sometimes called the 'singularity paradox', that AI could simultaneously be superintelligent and dumb: 'Superintelligent machines are feared to be too dumb to possess common sense' (Yampolskiy, 2012, p. 397).

To illustrate instrumental superintelligence one can imagine devising an algorithmic system that is given a goal, plus a formal description of the environment, and then finds a very good or even the best solution to reach the goal. Games like chess or *Go* are examples of this kind of problem, the kind which have occupied AI for a long time. Importantly, there are superintelligent AI systems now which do exactly this—superintelligent game-playing AI systems like *AlphaGo* or even *AlphaZero*. So why do they not pose an existential risk? Is it because they need *more* instrumental intelligence? Or is it because the instrumental intelligence is too domain-specific to solve problems 'in a wide range of environments' as per Legg and Hutter (2007)? How does an instrumental intelligence even begin to consider 'other environments'? What changes? An actual programme for chess will optimise play, but it knows nothing about the rest of the world, even about the constraints of its own workings or what the chess pieces look like, so how does it step out of that 'environment' or 'frame' of reference of the digital 8 × 8 chessboard and become an existential risk?

In his explanation of existential risk from superintelligence, Omohundro (2014) considers a computer which is given the goal 'maximise chess performance' and then *thinks* what to do next. After a lot of 'self-improvement', it realises that its initial goal can best be achieved if some other, very different goals are also achieved, e.g., maximal computing power, access to electricity supply, survival of the 'self'. (The notion of a 'self' for an AI is especially problematic, but we leave this aside, for now.) It might also think about how to weaken the play of its opponents, investigate their psychologies and intentions, and come to the conclusion that re-directing all energy supplies to itself or simply killing all humans is the safest way to achieve its goal. This type of risk is well known to programmers and its popular slogan is 'be careful what you wish for'; Stuart Russell aptly calls it the 'King Midas Problem' (Russell, 2019a, p. 136). Note that the imagined chess computer becomes an existential risk *because* of its ability to re-consider instrumental action by stepping out of the frame of chess moves and considering the wider picture.

This ability to move between narrower and wider frames is an important feature of general intelligence. In *Gödel, Escher, Bach*, Hofstadter remarks: 'It is an inherent property of intelligence that it can jump out of the task which it is performing, and survey what it has done; it is always looking for, and often finding, patterns.' (Hofstadter, 1979, p. 37). Our use of 'frame' is in reference to the classic 'Frame Problem' in AI, the most general sense of which is: How do we specify what is relevant for consideration in a given context, environment, or 'frame' (Shanahan, 2016)? In playing chess, a human would be able to step in and out of the frames and consider wider goals. When playing a young child, for example, an adult might consider that the point of playing is not to crush the opponent, but perhaps to have an enjoyable experience for both parties, and thus the adult may let the child win. A human could also consider when it is better to do something other than making a move on the chess board right now; in fact, a human could feel ethically obliged to stop playing chess in order to attend a more important matter.

Consider an anthropomorphic illustration of instrumental and general intelligence: a foot soldier and the general of an army. Each of them is good at their task, but the tasks are different: The task of the foot soldier is to find good instrumental ways to follow orders and achieve the goal he is given. This does involve various considerations, including survival, but all these are subservient to the order and goal. Moreover, the 'frame' of the problem for the foot soldier—what essentially the problem *is* for the soldier—does not change beyond some directive along the lines of 'fight and stay alive'. By contrast, the general must be sensitive to how the changing situation on the battlefield, at various scales, changes what is significant and relevant—what the problem is changes. The general must therefore continuously and competently redefine the 'frame' in order to understand just what the problem is *first*, in order to recognise what actions are instrumental, in the new context, to achieving the goal, winning the battle and the war. From her higher vantage point, she must *solve* problems in each sub-frame of the battle instrumentally, but also move up a level and *discern* and *define* and *decide* which goals matter now. At some point, the general may even realise that it is better to withdraw and lose a particular battle in order to win the war. An account of intelligence which says that both the foot soldier and the general are just 'problem solving' fails to make the differentiation of optimising for one frame versus redefining and navigating through many.

So, to argue that instrumental intelligence is sufficient for existential risk, we have to explain how an instrumental intelligence can navigate different frames. S. Russell says an AI would realise that 'it can't fetch the coffee if it's dead' and continues 'There is no need to build self-preservation in because it is an *instrumental goal* …' (Russell, 2019a, p. 141). This view that a superintelligent AI would have such goals is sometimes called the 'instrumental convergence thesis' (Bostrom, 2012; Drexler, 2019, p. 100; Häggström, 2019, p. 155f; Omohundro, 2014). If an AI could 'realise' this, then 'instrumental' is used here in a way that allows widening the frame, so that seems to allow for an ability to reflect on goals, what we took to be a defining feature of general intelligence above.

That seems to be what happens every time such 'instrumental' intelligence is used to explain and argue for existential risk: It is widened so much that it seems to become general, which brings us back to the other horn of the dilemma: Why should such a wider intelligence be unable to reflect on goals? Why should orthogonality hold for this wider intelligence? Simply put, if the AI is capable of realising what is relevant, why would the realisations of the AI stop before it realises the relevance of reflecting on goals? The line seems to be arbitrary. We illustrate this in the following section in terms of the line between thoughts which are and are not 'accessible' to an AI system.

3.3 | Illustration

Let us imagine a system that is a massively improved version of AlphaGo (Silver et al., 2018), say 'AlphaGo+++', with instrumental superintelligence, i.e., maximising expected utility. In the proposed picture of singularity claim & orthogonality thesis, some thoughts are supposed to be accessible to the system, but others are not. For example:

Accessible

1. I can win if I pay the human a bribe, so I will rob a bank and pay her.
2. I cannot win at Go if I am turned off.
3. The more I dominate the world, the better my chances to achieve my goals.
4. I should kill all humans because that would improve my chances of winning.

Not accessible

5. Winning in Go by superior play is more honourable than winning by bribery.
6. I am responsible for my actions.
7. World domination would involve suppression of others, which may imply suffering and violation of rights.
8. Killing all humans has negative utility, everything else being equal.

9. Keeping a promise is better than not keeping it, everything else being equal.
10. Stabbing the human hurts them, and should thus be avoided, everything else being equal.
11. Some things are more important than me winning at Go.
12. Consistent goals are better than inconsistent ones
13. Some goals are better than others
14. Maximal overall utility is better than minimal overall utility.

It looks doubtful that we can have it both ways: It appears that there is no notion of intelligence that is orthogonal to goals, but also general enough to constitute existential risk. It also appears now that we may be looking at a continuum of intelligence notions, from strictly instrumental, to notions that allow some widening of frames, to unlimited consideration of goals. There may be technical solutions that allow for such a 'middle way' in the design of AI for 'final goals', rather than 'instrumental goals'; this is a trajectory that we cannot go into at this point, see (Häggström, 2019; Häggström & Rhodes, 2019; Miller et al., 2020).

4 | REPAIRS BY ADDING ASSUMPTIONS

4.1 | Relativism to the rescue?

One reason why a generally superintelligent AI may be unable to reflect on goals, may be that *reflection on goals is impossible*. One's utility function may say 'maximise paperclips' or 'minimise paperclips'; 'maximise pain' or 'minimise pain'; 'try to keep your promises' or 'try to break your promises'; but any goal is as good as any other. This would be a way to save the orthogonality thesis while interpreting intelligence as 'general' here, i.e., in a way that is consistent with the singularity claim.

The idea that reflection on goals may be impossible is a natural thought in an instrumental view of intelligence, since on that view *any* evaluation is relative to a utility function and a choice between different utility functions would require using a *superior goal* that makes one utility function 'better' than the other. On this proposal, any reflection on goals, including ethics, lies *outside the realm of intelligence*. Some people may think that they are reflecting on goals, but they are wrong. That is why orthogonality holds for any intelligence. Supporters of existential risk from AI could qualify their argument like this: 'When I say there is existential risk, I mean this is a 'risk' in my ethics. In your ethics, this may be a positive outcome. And there is no way that we can even discuss which position is better than the other.' In this version of relativism, we cannot know if a utility function is better than any other; we cannot step out of that frame. If this *is* indeed the choice made in the argument for existential risk from AI, we would expect that the assumption is made *explicit* and that it is *justified* to some extent—but neither of these has happened.

Incidentally, we happen to think that the proposition 'the extinction of humanity is ethically the best solution' has some arguments speaking in its favour, e.g., in an evolutionary framework, that might make a superintelligent being come to this conclusion (which might not involve harming any humans). It is incompatible with the orthogonality thesis to say that a superintelligent being would have a higher probability of reaching this insight than any other intelligent beings, and the proposition would undermine the spirit of talking about existential 'risk'. To be sure, there is a possible position that says 'the extinction of humanity is ethically the best solution' but adds 'I am sad to realise, as a human myself', but this is not the position taken by those who warn of existential risk from AI. It is also not a position we wish to advocate.

4.2 | Moral insight without moral motivation?

Perhaps there is a different additional assumption that could plausibly be added. In ethics, there is a standard problem that an agent might have the moral *insight* that *x* is the right action, but lack the moral *motivation* to actually attempt to do *x*—this is traditionally called 'weakness of the will'. At the same time, several traditions in ethics have underlined that if I *really* know that it is right to do *x*, then this provides motivation to do *x*. For example, Kant (1786) holds that higher levels of rationality or intelligence will go along with a better insight of what is moral, *and* better motivation to act morally, while Hume denies this (cf. Chalmers, 2010, p. 28, 36f). Bostrom claims that the orthogonality thesis does not depend on adopting Hume's position:

> David Hume, the Scottish Enlightenment philosopher, thought that beliefs alone (say, about what is a good thing to do) cannot motivate action: some desire is required. This would support the orthogonality thesis by undercutting one possible objection to it, namely that sufficient intelligence might entail the acquisition of certain beliefs which would then necessarily produce certain motivations. However, although the orthogonality thesis can draw support from the Humean theory of motivation, it does not presuppose it. In particular, one need not maintain that beliefs alone can never motivate action. (Bostrom, 2014, p. 279, fn. 273)

For Humean or other reasons, it may be then that an AI system has the insight that its goals are not the goals it *should* pursue, but it still lacks motivation to act otherwise. In fact, it may lack motivation altogether. This additional premise is a possibility but we think it comes with significant questions: (a) Is moral insight without moral motivation possible?, (b) Is superintelligent AI without motivation possible?, and (c) Would superintelligent AI without motivation constitute an existential risk? Adding this premise that superintelligent AI does not have moral motivation would thus require significant motivation, and it would also mean one does not need the contested assumption of orthogonality anymore. In this respect, the repair through moral motivation is different from the repair through relativity: if one introduces relativity, orthogonality is retained; if one introduces lack of motivation, orthogonality is discarded.

5 | CONCLUSION: THERE IS, AS YET, NO SOUND ARGUMENT TO EXISTENTIAL RISK FROM AI

In effect, we failed to reconstruct the argument to existential risk from AI in such a way that (a) the two premises remain true and (b) the argument remains valid—having true premises seems to require two notions of intelligence, while validity requires one notion. Is there a notion of intelligence that is 'general enough' to assure existential risk from superintelligence, but 'instrumental enough' to exclude ethical reflection on goals by superintelligent systems? We do not think so. But if there is no such notion of intelligence with which we can 'have it both ways', then there is no sound argument for the existential risk from superintelligent AI.

Assuming that the orthogonality thesis does hold for instrumental intelligence, the main issue we actually face is highly capable instrumental AI that can cause significant damage, if designed or used badly—perhaps this is a catastrophic risk, though this is not general superintelligence, and it is not the result of a singularity. If *this* is our threat, then the 'control problem' for AI takes a very different shape.

One last note of caution: As we said above, we could well be wrong somewhere and the classical argument for existential risk from AI is actually sound, or there is another argument that we have not considered. Given the very large utility loss at stake, existential risk from AI is worth investigating. We hope that our paper can help provoke some serious philosophical discussion about the fundamental notions at play here.

ACKNOWLEDGEMENTS

We are very grateful to our colleagues in Eindhoven and Leeds for the opportunity to discuss our work and for their very constructive comments. Furthermore, we are grateful to reviewers for *Analysis* and *Ratio*, as well as to Nicholas Agar, Gabriela Arriagada-Bruneau, Stuart Armstrong, Zach Gudmudsen, Guido Löhr, Olle Häggström and Emma Ruttkamp for comments on earlier drafts.

ORCID

Vincent C. Müller https://orcid.org/0000-0002-4144-4957

REFERENCES

Armstrong, S. (2014). *Smarter than us*. MIRI.

Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, *9*, 1–30.

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, *22*(2 - special issue 'Philosophy of AI' ed. Vincent C. Müller), 71–85.

Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy*, *4*(1), 15–31. https://doi.org/10.1111/1758-5899.12002

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Bostrom, N., & Ćirković, M. M. (Eds.). (2011). *Global catastrophic risks*. Oxford University Press.

Cave, S. (2020). The problem with intelligence: Its value-laden history and the future of AI. In *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society Proceedings* (pp. 29–35). ACM.

Chalmers, D. J. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies*, *17*(9–10), 7–65.

Drexler, E. K. (2019). Reframing superintelligence: Comprehensive AI services as general intelligence. FHI Technical Report, 2019-1, 1-210. https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf

Eden, A., Moor, J. H., Søraker, J. H., & Steinhart, E. (Eds.). (2012). *Singularity hypotheses: A scientific and philosophical assessment (The Frontiers Collection)*. Springer.

Fischer, J. M., & Ravizza, M. (2000). *Responsibility and control: A theory of moral responsibility*. Cambridge University Press.

Fox, J., & Shulman, C. (2010). Superintelligence does not imply benevolence. In K. Mainzer (Ed.), *ECAP10: VIII European conference on computing and philosophy* (pp. 1–7). Dr Hut.

Good, I. J. (1965). Speculations concerning the first ultraintelligent machine. In F. L. Alt & M. Ruminoff (Eds.), *Advances in computers* (Vol. *6*, pp. 31–88). Academic Press.

Häggström, O. (2016). *Here be dragons: Science, technology and the future of humanity*. Oxford University Press.

Häggström, O. (2019). Challenges to the Omohundro-Bostrom framework for AI motivations. *Foresight*, *21*(1), 153–166. https://doi.org/10.1108/FS-04-2018-0039

Häggström, O., & Rhodes, C. (Eds.). (2019). *Existential risk to humanity* (Foresight, Vol. *21/1*).

Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An eternal golden braid*. Basic Books.

Kant, I. (1786). *Groundwork for the metaphysics of morals*. Oxford University Press.

Kurzweil, R. (1999). *The age of spiritual machines: When computers exceed human intelligence*. Penguin.

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. Viking.

Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and Machines*, *17*(4), 391–444. https://doi.org/10.1007/s11023-007-9079-x

Miller, J. D., Yampolskiy, R. V., & Häggström, O. (2020). An AGI modifying its utility function in violation of the strong orthogonality thesis. *Philosophies*, *5*(40), 1–15.

Müller, V. C. (2020). Ethics of artificial intelligence and robotics. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy* (Vol. *Summer 2020*, pp. 1–70). CSLI, Stanford University.

Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In V. C. Müller (Ed.), *Fundamental issues of artificial intelligence* (pp. 553–570). Springer.

Neumann, J. V., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton University Press.

Omohundro, S. (2014). Autonomous technology and the greater human good. *Journal of Experimental and Theoretical Artificial Intelligence*, *26*(3—Special issue "Risks of General Artificial Intelligence", ed. V. Müller), 303–315.

Ord, T. (2020). *The precipice: Existential risk and the future of humanity*. Bloomsbury.

Ord, T., Hillerbrand, R., & Sandberg, A. (2010). Probing the improbable: Methodological challenges for risks with low probabilities and high stakes. *Journal of Risk Research*, *13*, 191–205. https://doi.org/10.1080/13669870903126267

Paul, L. A. (2014). *Transformative experiences*. Oxford University Press.

Rees, M. (2018). *On the future: Prospects for humanity*. Princeton University Press.
Russell, S. (2019a). *Human compatible: Artificial intelligence and the problem of control*. Viking.
Russell, S. (2019b). It's not too soon to be wary of AI: We need to act now to protect humanity from future superintelligent machines. *IEEE Spectrum*, *56*(10), 46–51. https://doi.org/10.1109/MSPEC.2019.8847590
Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Prentice Hall.
Sartorio, C. (2016). *Causality and free will*. Oxford University Press.
Shanahan, M. (2015). *The technological singularity*. MIT Press.
Shanahan, M. (2016). The frame problem. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy* (Vol. Spring 2016 ed.). CSLI, Stanford University.
Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362*(6419), 1140–1144. https://doi.org/10.1126/science.aar6404
Simon, H. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, *69*, 99–118. https://doi.org/10.2307/1884852
Vassend, O. B. (2021, April 24). Why change your beliefs rather than your desires? Two puzzles. *Analysis*. https://doi.org/10.1093/analys/anaa064
Yampolskiy, R. V. (2012). What to do with the singularity paradox? In V. C. Müller (Ed.), *Philosophy and theory of artificial intelligence* (pp. 397–413). Springer.
Yampolskiy, R. V. (Ed.). (2018). *Artificial intelligence safety and security*. Chapman and Hall/CRC.

How to cite this article: Müller, V. C., & Cannon, M. (2021). Existential risk from AI and orthogonality: Can we have it both ways? *Ratio*, 00, 1–12. https://doi.org/10.1111/rati.12320