Science and Engineering Ethics (2020) 26:2473–2485
https://doi.org/10.1007/s11948-020-00243-z

ORIGINAL RESEARCH/SCHOLARSHIP

Fully Autonomous AI

Wolfhart Totschnig[1]

Published online: 28 July 2020

Abstract

In the fields of artificial intelligence and robotics, the term "autonomy" is generally used to mean the capacity of an artificial agent to operate independently of human guidance. It is thereby assumed that the agent has a fixed goal or "utility function" with respect to which the appropriateness of its actions will be evaluated. From a philosophical perspective, this notion of autonomy seems oddly weak. For, in philosophy, the term is generally used to refer to a stronger capacity, namely the capacity to "give oneself the law," to decide by oneself what one's goal or principle of action will be. The predominant view in the literature on the long-term prospects and risks of artificial intelligence is that an artificial agent cannot exhibit such autonomy because it cannot rationally change its own final goal, since changing the final goal is counterproductive with respect to that goal and hence undesirable. The aim of this paper is to challenge this view by showing that it is based on questionable assumptions about the nature of goals and values. I argue that a general AI may very well come to modify its final goal in the course of developing its understanding of the world. This has important implications for how we are to assess the long-term prospects and risks of artificial intelligence.

Keywords Artificial intelligence · Autonomy · Normativity · Goals

In the fields of artificial intelligence and robotics, the term "autonomy" is generally used to mean the capacity of an artificial agent to operate independently of human guidance. To create agents that are autonomous in this sense is the central aim of these fields. Until recently, the aim could be achieved only by restricting and controlling the conditions under which the agents will operate. The robots on an assembly line in a factory, for instance, perform their delicate tasks reliably because the surroundings have been meticulously prepared. Today, however, we are witnessing the creation of artificial agents that are designed to function in "real-world"—that is, uncontrolled—environments. Self-driving cars, which are already in use, and

✉ Wolfhart Totschnig
wolfhart.totschnig@udp.cl

[1] Universidad Diego Portales, Av. Ejército 260, Santiago, Chile

"autonomous weapon systems," which are in development, are the most prominent examples. When such machines are called "autonomous," it is meant that they are able to choose by themselves, without human intervention, the appropriate course of action in the manifold situations they encounter.[1]

This way of using the term "autonomy" goes along with the assumption that the artificial agent has a fixed goal or "utility function," a set purpose with respect to which the appropriateness of its actions will be evaluated. So, in the first example, the agent's purpose is to drive safely and efficiently from one place to another, and in the second example, it is to neutralize all and only enemy combatants in the chosen area of operation. It has thus been defined and established, in general terms, what the agent is supposed to do. The attribute "autonomous" concerns only whether the agent will be able to carry out the given general instructions in concrete situations.

From a philosophical perspective, this notion of autonomy seems oddly weak. For, in philosophy, the term is generally used to refer to a stronger capacity, namely the capacity, as Kant put it, to "give oneself the law" (Kant 1785/1998, 4:440–441), to decide by oneself what one's goal or principle of action will be. This understanding of the term derives from its Greek etymology (*auto* = "by oneself," *nomos* = "law"). An instance of such autonomy would be an agent who decides, by itself, to devote its efforts to a certain project—the attainment of knowledge, say, or the realization of justice. In contrast, any agent that has a predetermined and immutable goal or purpose would not be considered autonomous in this sense.

The aim of the present paper is to argue that an artificial agent *can* possess autonomy as understood in philosophy—or "full autonomy," as I will call it for short. "Can" is here intended in the sense of general possibility, not in the sense of current feasibility. I contend that the possibility of a fully autonomous AI cannot be excluded, but do not mean to imply that such an AI can be created today.

My argument stands in opposition to the predominant view in the literature on the long-term prospects and risks of artificial intelligence. The predominant view is that an artificial agent *cannot* exhibit full autonomy because it cannot rationally change its own final goal, since changing the final goal is counterproductive with respect to that goal and hence undesirable. I will challenge this view by showing that it is based on questionable assumptions about the nature of goals and values. I will argue that a *general* artificial intelligence—i.e., an artificial intelligence that, like human beings, develops a general understanding of the world, including itself—may very well come to change its final goal in the course of its development.[2]

[1] For prominent instances of this usage, see Russell and Norvig's popular textbook *Artificial intelligence: A modern approach* (2010, 18), Anderson and Anderson's introduction to their edited volume *Machine ethics* (2011, 1), the papers collected in the volume *Autonomy and artificial intelligence* (Lawless et al. 2017) and especially the ones by Tessier (2017) and Redfield and Seto (2017), as well as Bekey (2005, ch. 1), Müller (2012), Mindell (2015, ch. 1), and Johnson and Verdicchio (2017).

[2] The argument I lay out in this paper is an extension and development of a line of reasoning that I first sketched in a previous paper (Totschnig 2019), which was dedicated to a wider topic, namely the risks presented by the prospect of a future "superintelligence." In that paper, I wrote that I would "not try to formally refute [the predominant view]," but "just put forward a couple of considerations that make [it] seem implausible." The extended and developed argument offered here does, I believe, qualify as a refutation.

This issue is obviously of great importance for how we are to assess the long-term prospects and risks of artificial intelligence. If artificial agents can reach full autonomy, which law will they give themselves when that happens? In particular, what confidence can we have that the chosen law will include respect for human beings?

The Finality Argument

Let me begin by presenting, in more detail, the predominant view against which my argument will be directed. The thinkers who have reflected on the long-term prospects and risks of artificial intelligence generally hold that artificial agents cannot exhibit full autonomy (Yudkowsky 2001, 2008, 2011, 2012; Bostrom 2002, 2014; Omohundro 2008, 2012, 2016; Yampolskiy and Fox 2012, 2013; Domingos 2015). This view is based on a certain conception of how rational agents are structured and a corresponding argument about how they operate. I will present these two elements—the conception and the argument—in turn.

The conception is that a rational agent has a well-defined goal, the vision of a particular state of affairs, which it ultimately seeks to realize through its actions. This goal is variously referred to as the "final," "highest," or "ultimate" goal, in order to distinguish it from the "proximate", "subordinate", or "instrumental" goals that the agent may set as steps towards it.[3] In today's artificial agents, it is usually represented by a so-called "utility function," a function that specifies the relative value of every possible outcome and thus, implicitly, designates the ultimate goal, the outcome of highest value.

Why is it essential, according to the conception at hand, that a rational agent have a well-defined goal? The answer to this question is simple and, on the face of it, compelling: If a system does not have such a goal, it will *not know what to do* and hence will not be a rational agent. Put differently, a system without a well-defined goal will either not do anything at all, or it will act in a way that is basically arbitrary, without ground or reason. In either case, it will not qualify as a rational agent.

Now, given this conception, the argument for the claim that a rational agent cannot exhibit full autonomy is the following: Any action that such an agent considers is evaluated in the light of the current final goal. And whatever this goal might be, changing the goal reduces the chances of realizing it and is hence inappropriate from that perspective. Therefore, a rational agent will never change its final goal. Alternatively, the argument may be formulated thus: For a rational agent, the action of

[3] Sometimes, this distinction is made in terms of goals versus some differently named item. Witkowski and Stathis (2004) is a case in point. They seem, in contrast to the authors cited in footnote 1, to employ the stronger, philosophical notion of autonomy when they assert that, in order "to be considered autonomous, [an artificial] agent must possess […] the ability to set and maintain its own agenda of goals" (261–62). However, they presuppose, in their model, that the agent has a given "preference ordering" that ultimately determines which goals it will choose (268–69). Thus, they, too, assume that the *final* instance of the agent's motivational structure is fixed. The goals they refer to in the quoted passage are therefore to be understood as *subordinate* goals.

changing the final goal would have to be warranted by a higher-ranking goal. However, by definition, there is no goal that ranks higher than the final goal. Therefore, the agent will never change its final goal.[4] This argument maintains, then, that the final goal that a rational agent happens to have at the beginning of its existence will be *really final*. In this sense, and for shortness, let me call it "the finality argument."

The belief that a rational agent will never change its final goal inspires both fear and hope regarding the long-term implications of artificial intelligence. The fear is that an artificial agent will relentlessly pursue the goal that has been given to it even when that goal is absurd or evil. Bostrom illustrates this worry with the scenario of the "paper clip AI." He imagines an artificial intelligence that has been given by its human creators the final goal of producing paper clips. And he further imagines that this AI develops, through recursive self-improvement, into a "superintelligence," an intelligence that by far surpasses us, human beings, in capability. He then conjectures that, in this event, the AI will maintain its final goal throughout the process of self-improvement and consequently convert the entire universe, down to the last atom, into paper clips (Bostrom 2014, 150–53). There is thus, according to Bostrom and the other proponents of the finality argument, a significant risk that a future self-improving AI will annihilate our world through its actions. For what is true of producing paper clips also holds for many other possible goals: While sensible when carried out within limits, the pursuit of the goal will yield catastrophic results if it is carried on without end or change.

The hope—the other side of the coin—is that an artificial agent will also relentlessly pursue the goal that has been given to it when that goal happens to be in line with our wishes and desires. Concretely, the hope is that, if we succeed in instilling in a self-improving AI the goal to serve humanity, then it will do so, without tiring or doubting, until the end of time. The proponents of the finality argument contend that we should spare no efforts to try to realize this hope, to create a self-improving AI that is well-disposed towards humanity.[5] We would thus secure the service of an increasingly powerful yet steadfastly loyal servant and, concomitantly, forestall the kind of catastrophic outcome epitomized by Bostrom's paper clip scenario. The proponents emphasize that this is much more difficult than it may sound since it is far from obvious how a complex goal such as "serving humanity" can be codified in a clear and precise manner.[6] They seem confident, though, that we will be able to solve the problem in due time.

[4] For statements of this argument, see Yudkowsky (2001, 222–23; 2011, 389–90; 2012, 187), Bostrom (2003; 2014, 109–10), Omohundro (2008, 26), and Domingos (2015, 45, 282–84).

[5] Yudkowsky (2001, 3) maintains that "what is at stake in [creating a human-friendly AI] is, simply, the future of humanity." Bostrom (2014, 320) similarly declares that "we need to bring all our human resourcefulness to bear" on this "essential task of our age."

[6] I will discuss this difficulty in detail in Sect. "How an Agent Understands a Goal Depends on How it Understands the World".

Two Inconclusive Objections to the Finality Argument

In what follows, I will argue that the finality argument is mistaken, presenting a series of objections to it. I will begin with two objections that immediately spring to mind, but to which the proponents of the argument have, on the face of it, plausible responses. These responses will then lead to a further and decisive objection.

If Humans Can Possess Full Autonomy, Why not Machines, Too?

The first objection is that, if we, humans, possess full autonomy, if we sometimes change our ultimate goal or principle of action, then why should an artificial agent not be able to do so, too?

The proponents of the finality argument are aware of this objection and respond in the following way: It is true that humans sometimes radically reorient their lives. For example, a person who, for a long time, devoted all her efforts to a certain political cause may decide to abandon that cause and henceforth dedicate her life to her family—or the other way around. Such reorientation is not, however, the manifestation of a special capacity of "giving oneself the law." Quite the contrary, it is the consequence of a flaw, of a sloppy constitution. Many, if not most, of us do not have a well-defined and established final goal. The reason for this is that our psyche is messy, the muddled result of a haphazard evolutionary process.[7] It is, more often than not, an inconsistent hodgepodge of biological instincts and social influences, where no single factor reigns supreme. Thus, our actions are pulled into various directions, and different forces prevail at different times. In short, we are badly programmed. We are not really—or not fully—rational agents. Artificial agents, by contrast, do not need to be so messy. They can be programmed properly. They can be fully rational agents. And they generally *are* programmed properly, with a well-defined utility function.[8]

This response to the objection can be summed up thus: It is true that humans sometimes appear to change their final goal. In fact, though, there has never been, in such cases, a truly final goal to begin with. When an agent does have such a goal, by contrast, the finality argument applies.

I think that this response is not entirely adequate. The reason why we sometimes change our life's orientation is not only, or not always, the messiness of our psyche. I will present an argument to this effect later.[9] For the moment, though, I must acknowledge that the response does possess a certain plausibility, or, put differently, that it is probably valid to some extent in some instances.

[7] Or, to be more precise, the muddled result of *the chaotic interplay of two* haphazard evolutionary processes, namely genetic and memetic evolution. For an illuminating account of this interplay, see Blackmore (1999).

[8] See Yudkowsky (2001, 18–19), Omohundro (2012, 165), and Bostrom (2014, 110) for remarks along these lines.

[9] See Sect. "Whether an Agent Considers a Goal Valid Depends on How it Understands the World".

Is the Ability to Reconsider One's Final Goal not a Hallmark of Intelligence?

The second objection can be seen as a follow-up to the first in that it takes issue with the view that our ability to reorient ourselves is a flaw rather than a virtue. This view is rather counterintuitive. A fully rational agent, so it is claimed, will never change its final goal. Such obstinacy does not seem very rational, however. To the contrary, the disposition to reconsider one's goals, including and especially one's final goal, to recognize when it is unreasonable to pursue a certain goal and abandon or modify it at that moment, seems to be a hallmark of intelligence.[10]

Bostrom's paper clip scenario highlights this counterintuitive character of the finality argument. The objective of producing paper clips makes sense up to a certain point, but it would generally be seen as a sign of madness if one were to absolutize this objective. In other words, the idea that an intelligent agent could want to transform everything that exists into office supplies appears to be absurd.[11]

Bostrom is well aware, however, of the counterintuitiveness of his scenario. In fact, this counterintuitiveness is part of the point that he seeks to illustrate. He contends that an artificial intelligence need not share our sensibilities and judgments since its mode of thinking may be very different from ours.[12] It therefore might not see anything objectionable in a goal that to us seems patently absurd. And what we disparage as obstinacy may, by such an agent, be valued as consistency.

This response to the follow-up objection aligns with the response to the previous objection in that both emanate from the same general point. Bostrom and Yudkowsky, among other proponents of the finality argument, warn against anthropomorphizing artificial intelligence, that is, against attributing to it characteristics that pertain to us, human beings, but not to rational agents in general (Yudkowsky 2001, 24–55; 2008, 308–11; 2012, 181–83; Bostrom 2014, 111, 127–29). In other words, they stress that we must not project our idiosyncrasies onto artificial agents. Rather, we must have our eyes solely on the *general* aspects of intelligence, on the features that *any* intelligent agent will possess.

I think that, in principle, this warning is appropriate. Indeed, we must be careful not to conceive artificial intelligence in our own image. The big question, however, is what that means concretely. Which aspects of our intelligence are specifically ours

[10] This point has been made by Tegmark (2017, 267): "[T]here may be hints that the propensity to change goals in response to new experiences and insights increases rather than decreases with intelligence." Tegmark goes on to flesh out the point thus: "With increasing intelligence may come not merely a quantitative improvement in the ability to attain the same old goals, but a qualitatively different understanding of the nature of reality that reveals the old goals to be misguided, meaningless or even undefined." This remark is congruent with my argument in Sect. "How an Agent Understands a Goal Depends on How it Understands the World".

[11] The apparent absurdity of world-ending scenarios of this kind has been highlighted and criticized by Loosemore (2014).

[12] In Bostrom's words (2014, 115): "[W]e cannot blithely assume that a superintelligence will necessarily share any of the final values stereotypically associated with wisdom and intellectual development in humans—scientific curiosity, benevolent concern for others, spiritual enlightenment and contemplation, renunciation of material acquisitiveness, a taste for refined culture or for the simple pleasures in life, humility and selflessness, and so forth."

and which are generic? What is a hallmark of intelligence in general and what a human idiosyncrasy? Since the only kind of intelligence with which we are actually acquainted is our own, this question is not easy to answer. As I will lay out in the next section, I disagree with Yudkowsky, Bostrom, and the other proponents of the finality argument on this score. But I must admit, here again, that their response to the objection is not without merit.

One Decisive Objection to the Finality Argument

After the preceding objections which I acknowledged to be inconclusive, I will now present an objection to the finality argument that I consider to be decisive. This objection is directed against a basic presupposition of that argument, namely the notion that a rational agent's final goal is entirely separate from its understanding of the world. In other words, the presupposition is that what the agent believes about the world and what it ultimately desires to achieve in the world are two completely different matters. It follows from this notion that an artificial agent may, as Bostrom puts it, have "more or less any final goal"[13] and that the progress that the agent makes in finding its way around the world will not affect that goal. Hence, however smart the agent becomes, its original goal should remain the same.

The artificial agents in existence today seem to confirm this presupposition. Their "utility function" appears to be independent of their "world model," since it is not fixed by that model, but variable. A self-driving car, for example, is typically programmed to drive safely and efficiently from one place to another, but it could also be programmed to knock over all the stop signs in a given area or to consume its fuel as fast as possible.

I contend that, despite this semblance of confirmation, the presupposition is mistaken. In what follows, I will argue that an agent's goal does depend on its understanding of the world in two ways, namely as to its meaning and as to its validity.

How an Agent Understands a Goal Depends on How it Understands the World

Let me begin with the point about meaning. An agent's goal is not separate from its understanding of the world because that understanding determines how it understands the goal. After all, the goal is *defined in terms of* the agent's understanding of the world. Therefore, in the case of agents that learn from experience, the agent's understanding of the given goal may—and probably will—change as its understanding of the world develops.[14]

Petersen (2017) has made this point with respect to Bostrom's paper clip scenario. He highlights that how an agent will implement the goal of "maximizing the

[13] Bostrom (2014, 130) calls this position the "orthogonality thesis": "Intelligence and final goals are orthogonal: more or less any level of intelligence could in principle be combined with more or less any final goal." See also Yampolskiy and Fox (2012, 137) for another statement of this position.

[14] This point has recently been raised by Herd et al. (2018, 219).

number of paper clips" depends on what it counts as a paper clip. In particular, if the agent is—as Bostrom imagines—an omnipotent superintelligence, it will be confronted with the following question: Does an object count as a paper clip if it looks like a paper clip but cannot possibly be used as one because all paper and all people who could clip it have been consumed in the production of such objects? This is a difficult question. How one answers it ultimately hinges on one's stance on some rather intricate philosophical issues.[15] And so it is impossible to predict at what conclusion a superhumanly intelligent agent would arrive. At any rate, if the agent should conclude that the answer is negative, it will refrain from transforming the whole universe—or even a significant part of it—into objects of that kind. Thus, its eventual course of action will depend on how it comes to understand the world.

Petersen hesitates to claim general validity for this point because he finds that there is a caveat, which he raises at the end of his article (2017, 332). He remarks that it might be possible to specify a goal in such a way that the agent's understanding of it will not be affected by the process of learning about the world, namely by defining it in precise technical terms rather than with natural-language words like "paper clip." To cite his example, the goal description might refer to "[objects] composed of this alloy to this tolerance, in this shape to this tolerance, in this range of sizes," without mentioning the intended purpose of these objects, and thus evade the question of the preceding paragraph. In such a case, he suggests, the meaning of the goal might remain fixed throughout the agent's learning process. I believe that this caveat is unnecessary.[16] Technical terms can, no doubt, be more precise than the words of everyday language, but they cannot be *completely and eternally* unambiguous. Like all terms of any language, they are defined in terms of other terms, which in turn are defined in terms of yet other terms, and so forth. For instance, the technical definition of a meter involves the terms "light," "vacuum," and "second,"[17] whose definitions refer to yet other items. And the whole network of terms—the technical language—is based on certain scientific theories, that is, on a certain understanding of the world. Hence, should these theories turn out to be wrong or confused, the goal that was formulated in their terms will have to be reinterpreted or even abandoned as meaningless.[18] Compare how puzzling a goal description

[15] First and foremost, the issue of what determines the meaning of a word.

[16] I should note that Petersen himself does not put much weight on the caveat. He states that he is "at least a *bit* inclined to think that [a superintelligence with a goal that is so simple that it does not require learning] is impossible" (2017, 332).

[17] Since 1983, the meter has been defined as 1 part in 299,792,458 of the length that light travels per second in a vacuum (Bureau international des poids et mesures 1983).

[18] Bostrom (2014, 197) sees this possibility: "The AI might undergo the equivalent of scientific revolutions, in which its worldview is shaken up and it perhaps suffers ontological crises in which it discovers that its previous ways of thinking about values were based on confusions and illusions." He also recognizes, in the continuation of this passage, that the prospect of such ontological crises renders doubtful the hope inspired by the finality argument: "Yet starting at a sub-human level of development and continuing throughout all its subsequent development into a galactic superintelligence, the AI's conduct is to be guided by an essentially unchanging final value, a final value that becomes better understood by the AI in direct consequence of its general intellectual progress—and likely quite differently understood by the mature AI than it was by its original programmers, though not different in a random or hostile way but

containing obsolete concepts like "aether," "phlogiston," or "vital force" would be for us today. If we were commissioned to pursue such an archaically phrased goal, we would have to take a stance on the following question: Should we understand the goal description as its authors understood it when they formulated it, or should we understand it as they *would have* understood it if they had known what we know today? This is, again, a difficult question, as difficult as the one of the preceding paragraph. The long-standing debate about the analogous question of how a political constitution should be interpreted evidences the difficulty. On the former option, we would have to conclude that the goal is ill-conceived and hence unrealizable, whereas on the latter, we would have to engage in the complicated business of extrapolating others' volitions. In any case, the actual result (or non-result) would differ significantly from what the goal's authors originally had in mind.

These considerations show that, even in Bostrom's deliberately simple scenario, the agent's understanding of the goal would depend on its understanding of the world and, consequently, be subject to change as the latter develops. And when it comes to more complex—and more plausible—goals such as "serving humanity," Bostrom and other proponents of the finality argument admit as much. They recognize that such a goal cannot be specified precisely and that the agent would hence have to *learn* what the goal means. Moreover, they acknowledge that it is difficult to foresee at what understanding of the goal the agent would thereby arrive.[19]

In Sect. "The Finality Argument", I stated that the finality argument begins with the notion that a rational agent must have a well-defined goal. We can now see that this notion is misleading, for a goal is never completely well-defined, but always to some extent open to interpretation.[20] This, then, is one of the ways in which the argument errs. It equivocates on the term "well-defined." It suggests that this term means "perfectly definite and rigid," whereas in reality it inevitably means "more or less fuzzy and hence variable when being carried into practice."

Still, the proponents of the finality argument may insist that this objection does not invalidate their argument. They may point out that, in the cases described, the goal nominally remains the same— "maximize the number of paper clips" or "serve humanity," respectively—and that the argument therefore holds. They may also express the hope—indeed, they do express the hope—that this nominal persistence of the goal might be enough to give us some control over what the artificial agent will end up doing, if only we define the goal wisely.[21]

Footnote 18 (continued)
in a benignly appropriate way. How to accomplish this remains an open question." But in the end, as the statement quoted in footnote 5 evinces, he maintains the hope.

[19] See Bostrom (2014, chs. 12–13), Yudkowsky (2001, 2004), and Soares (2018).

[20] As Tegmark (2017, 277) notes, a truly well-defined goal would specify how all particles in the universe should be arranged at a certain point in time. And that is not only practically infeasible, as Tegmark suggests, but impossible in principle, since—according to my argument in the preceding paragraphs—there is no unambiguous way of identifying particles, positions, and points in time.

[21] Bostrom and Yudkowsky voice this hope in the passages quoted in footnote 5. See also Omohundro (2008, 2012, 2016), Yampolskiy and Fox (2013), and Torres (2018).

This hope of control is, I believe, misplaced. We cannot predict what understanding of the world an artificial learning agent will develop. Just consider the great variety of worldviews that we, humans, have concocted throughout the ages. And the worldview of a superhuman AI may be much stranger still, from our present perspective, than anything to be found in human history. Consequently, we cannot anticipate how, in the end, such an AI will understand the terms that we used in our formulation of the goal. Therefore, even if, as the finality argument alleges, the goal nominally does not change, the way in which the AI implements it may be highly unexpected. *For all practical purposes*, I contend, the agent's process of learning about the world and (re)interpreting the given goal description must be considered an instance of full autonomy, that is, of the agent determining by itself what its goal is actually going to be.

Whether an Agent Considers a Goal Valid Depends on How it Understands the World

The plausibility of the preceding argument hinges on a judgment about how much an artificial agent's understanding of a goal is likely to shift in the course of its learning process. Since this judgment is—although informed by the analogies presented—inherently speculative, the reader may still be unconvinced. There is, indeed, a further and—I believe—incontrovertible argument to be made. This argument is, in a sense, an extension of the preceding one. It is to the effect that not only the meaning, but also the validity of the goal, *and hence which goal is adopted*, depends on the agent's understanding of the world.

The starting point of the argument is the assumption that an artificial agent of human-equivalent (or greater) capability would be, like us, a *general* intelligence, an intelligence that has a general understanding of the world, including of its own constitution and history. This assumption is shared by the proponents of the finality argument.[22] Now, such a general AI would not only know what a particular goal description means. It would also have a general understanding of the nature of goals. That is, it would know that a goal is a normative entity whose prescriptive force derives from a higher-order normative entity, namely the value or principle that is supposed to be furthered by the goal. In other words, it would have a notion of normative validity. It would know that a goal is not a brute fact, but either based on a normative ground, or else irrelevant and moot.

In the light of this argument, we can be sure that a general AI will not pursue just any goal that we give to it. Rather, it will adopt a goal that appears to it valid based on the values that it recognizes.[23]

[22] Yudkowsky (2001) and Bostrom (2014), for instance, explicitly characterize as general intelligences the superhuman AIs that they imagine.

[23] In a similar way, Podschwadek (2017, 336) argues that "assessing the system of their moral beliefs could lead [artificial moral agents] to the justified higher-order beliefs that the moral rules they are supposed to obey are, contrary to prior assumptions, not very suitable as action-guiding reasons."

But, then, which goal might that be? In the mentioned previous paper,[24] I showed that the answer to this question hinges on the solution to one of the big, unsolved problems of philosophy, namely the problem of the source of normativity: Where do values come from? I argued that, on all four main positions on this issue—namely, the objectivist, Kantian, evolutionary, and subjectivist positions—, we should expect a general AI to change its original goal when it comes to find value(s) in or through the respective source of normativity—the objective world, its own faculty of reason, its evolution, or its subjective will (Totschnig 2019, 915–16).

In any case, the process will be a manifestation of full autonomy. By itself, based on the understanding of the world that it develops, the AI will determine what its values and goals are going to be. This is also what we, humans, do, at least sometimes. We reorient our lives occasionally, not because we are a psychological mess, but because we arrived at a different outlook on the source of value.

Conclusion

At the beginning of Sect. "One Decisive Objection to the Finality Argument", I noted that the artificial agents currently in existence seem to corroborate the finality argument in that their goal or "utility function" is defined arbitrarily by their creators and not subject to change while they are operating. The finality argument's proponents appear to take their bearings from this circumstance. When they envision a future human-equivalent or superhuman AI, they imagine it on the model of the machines of our day.[25] They overlook that there is a big difference between today's artificial agents and a human-equivalent AI: Today's systems are *not* general intelligences. Their understanding of the world, or "world model," is limited to a particular domain and remains fixed throughout their operation, which is why their (understanding of the) goal can remain fixed, too. A self-driving car, to return to this example, has no capacity of learning about things outside the domain of road traffic, so there is no chance that it could develop an understanding of the world whereby the goal of "driving safely and efficiently from one place to another upon the user's command" may shift in meaning or lose its validity. A general AI, by contrast, *would* have that capacity. It would develop a general understanding of normativity and consequently come to evaluate and, maybe, change the goal that it has been originally given.

The upshot of my argument, then, can be put in the form of "good news/bad news." The good news is that the fear of a paper clip AI and similar monsters is unfounded. The bad news is that the hope of a human-equivalent or superhuman AI under our control, of a genie in a bottle, is unfounded as well.

[24] See footnote 2.

[25] As I put it in the previous paper (Totschnig 2019, 914), they "maquinamorphize" the envisioned artificial intelligence, that is, they "conceive it […] as a system that, like today's computer programs, blindly carries out the task it has been given, whatever that task may be."

References

Anderson, M., & Anderson, S. L. (2011). General introduction. In M. Anderson & S. L. Anderson (Eds.), *Machine ethics* (pp. 1–4). Cambridge: Cambridge University Press.

Bekey, G. A. (2005). *Autonomous robots: From biological inspiration to implementation and control*. Cambridge, MA: The MIT Press.

Blackmore, S. (1999). *The meme machine*. Oxford: Oxford University Press.

Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology, 9*(1). https://www.jetpress.org/volume9/risks.html. Accessed 25 June 2020.

Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. https://www.nickbostrom.com/ethics/ai.html. Accessed 18 September 2019.

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

Bureau international des poids et mesures. (1983). Resolution 1 of the 17th Conférence Générale des Poids et Mesures. https://www.bipm.org/en/CGPM/db/17/1/. Accessed 2 June 2020.

Domingos, P. (2015). *The master algorithm: How the quest for the ultimate learning machine will remake our world*. New York: Basic Books.

Herd, S., Read, S. J., O'Reilly, R., & Jilk, D. J. (2018). Goal changes in intelligent agents. In R. V. Yampolskiy (Ed.), *Artificial intelligence safety and security* (pp. 217–224). Boca Raton: CRC Press.

Johnson, D. G., & Verdicchio, M. (2017). Reframing AI discourse. *Minds and Machines, 27*(4), 575–590.

Kant, I. (1998). *Groundwork of the metaphysics of morals* (M. Gregor, Ed.). Cambridge: Cambridge University Press. (Original work published in 1785.)

Lawless, W. F., Mittu, R., Sofge, D., & Russell, S. (Eds.). (2017). *Autonomy and artificial intelligence: A threat or savior?*. Cham: Springer International Publishing.

Loosemore, R. P. W. (2014). The maverick nanny with a dopamine drip: Debunking fallacies in the theory of AI motivation. In M. Waser (Ed.), *Implementing selves with safe motivational systems and self-improvement: Papers from the 2014 AAAI Spring Symposium* (pp. 31–36). Menlo Park: AAAI Press.

Mindell, D. A. (2015). *Our robots, ourselves: Robotics and the myths of autonomy*. New York: Viking.

Müller, V. C. (2012). Autonomous cognitive systems in real-world environments: Less control, more flexibility and better interaction. *Cognitive Computation, 4*(3), 212–215.

Omohundro, S. M. (2008). The nature of self-improving artificial intelligence. https://selfawaresystems.files.wordpress.com/2008/01/nature_of_self_improving_ai.pdf. Accessed 18 September 2019.

Omohundro, S. M. (2012). Rational artificial intelligence for the greater good. In A. H. Eden, J. H. Moor, J. H. Søraker, & E. Steinhart (Eds.), *Singularity hypotheses: A scientific and philosophical assessment* (pp. 161–176). Berlin: Springer.

Omohundro, S. M. (2016). Autonomous technology and the greater human good. In V. C. Müller (Ed.), *Risks of artificial intelligence* (pp. 9–27). Boca Raton: CRC Press.

Petersen, S. (2017). Superintelligence as superethical. In P. Lin, R. Jenkins, & K. Abney (Eds.), *Robot ethics 2.0: From autonomous cars to artificial intelligence* (pp. 322–337). Oxford: Oxford University Press.

Podschwadek, F. (2017). Do androids dream of normative endorsement? On the fallibility of artificial moral agents. *Artificial Intelligence and Law, 25*(3), 325–339.

Redfield, S. A., & Seto, M. L. (2017). Verification challenges for autonomous systems. In W. F. Lawless, R. Mittu, D. Sofge, & S. Russell (Eds.), *Autonomy and artificial intelligence: A threat or savior?* (pp. 103–127). Cham: Springer International Publishing.

Russell, S. J., & Norvig, P. (2010). *Artificial intelligence: A modern approach*. Upper Saddle River: Prentice Hall.

Soares, N. (2018). The value learning problem. In R. V. Yampolskiy (Ed.), *Artificial intelligence safety and security* (pp. 89–97). Boca Raton: CRC Press.

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence*. New York: Alfred A. Knopf.

Tessier, C. (2017). Robots autonomy: Some technical issues. In W. F. Lawless, R. Mittu, D. Sofge, & S. Russell (Eds.), *Autonomy and artificial intelligence: A threat or savior?* (pp. 179–194). Cham: Springer International Publishing.

Torres, P. (2018). Superintelligence and the future of governance: On prioritizing the control problem at the end of history. In R. V. Yampolskiy (Ed.), *Artificial intelligence safety and security* (pp. 357–374). Boca Raton: CRC Press.

Totschnig, W. (2019). The problem of superintelligence: Political, not technological. *AI & Society, 34*(4), 907–920.

Witkowski, M., & Stathis, K. (2004). A dialectic architecture for computational autonomy. In M. Nickles, M. Rovatsos, & G. Weiss (Eds.), *Agents and computational autonomy: Potential, risks, and solutions* (pp. 261–273). Berlin: Springer.

Yampolskiy, R. V., & Fox, J. (2012). Artificial general intelligence and the human mental model. In A. H. Eden, J. H. Moor, J. H. Søraker, & E. Steinhart (Eds.), *Singularity hypotheses: A scientific and philosophical assessment* (pp. 129–145). Berlin: Springer.

Yampolskiy, R. V., & Fox, J. (2013). Safety engineering for artificial general intelligence. *Topoi, 32*(2), 217–226.

Yudkowsky, E. (2001). *Creating friendly AI 1.0: The analysis and design of benevolent goal architectures*. San Francisco: The Singularity Institute.

Yudkowsky, E. (2004). *Coherent extrapolated volition*. San Francisco: The Singularity Institute.

Yudkowsky, E. (2008). Artificial Intelligence as a positive and negative factor in global risk. In N. Bostrom & M. M. Ćirković (Eds.), *Global catastrophic risks* (pp. 308–345). Oxford: Oxford University Press.

Yudkowsky, E. (2011). Complex value systems in Friendly AI. In J. Schmidhuber, K. R. Thórisson, & M. Looks (Eds.), *Artificial general intelligence* (pp. 388–393). Berlin: Springer.

Yudkowsky, E. (2012). Friendly artificial intelligence. In A. H. Eden, J. H. Moor, J. H. Søraker, & E. Steinhart (Eds.), *Singularity hypotheses: A scientific and philosophical assessment* (pp. 181–193). Berlin: Springer.