Fully autonomous AI

Wolfhart Totschnig[0000−0003−2918−6286]

Universidad Diego Portales, Santiago, Chile

Note: This is an extended abstract of a paper that has been accepted for publication in *Science and Engineering Ethics*.

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 "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

[1] For prominent instances of this usage, see Russell & Norvig's popular textbook *Artificial intelligence: A modern approach* (2010, 18), Anderson & 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 one by Tessier (2017), as well as Müller (2012), Mindell (2015, ch. 1), and Johnson & Verdicchio (2017).

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 (Yudkowsky 2001, 2008, 2011, 2012; Bostrom 2002, 2014; Omohundro 2008, 2012, 2016; Yampolskiy & Fox 2012, 2013; Domingos 2015). 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 and of itself—may very well come to change its final goal in the course of its development.

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?

References

1. Anderson, M., Anderson, S.L.: General introduction. In: Anderson, M., Anderson, S.L. (eds.) Machine ethics, pp. 1–4. Cambridge University Press (2011)
2. Bostrom, N.: Existential risks. Journal of Evolution and Technology **9**(1) (2002), http://www.jetpress.org/volume9/risks.html
3. Bostrom, N.: Superintelligence. Oxford University Press (2014)
4. Domingos, P.: The master algorithm. Basic Books (2015)
5. Johnson, D.G., Verdicchio, M.: Reframing AI discourse. Minds and Machines **27**(4), 575–590 (2017)
6. Kant, I.: Groundwork of the metaphysics of morals. Cambridge Texts in the History of Philosophy, Cambridge University Press (1998)
7. Lawless, W.F., Mittu, R., Sofge, D., Russell, S. (eds.): Autonomy and artificial intelligence. Springer International Publishing (2017)
8. Mindell, D.A.: Our robots, ourselves. Viking (2015)
9. Müller, V.C.: Autonomous cognitive systems in real-world environments. Cognitive Computation **4**(3), 212–215 (2012)
10. Omohundro, S.M.: The nature of self-improving artificial intelligence (2008)
11. Omohundro, S.M.: Rational artificial intelligence for the greater good. In: Eden, A.H., Moor, J.H., Søraker, J.H., Steinhart, E. (eds.) Singularity hypotheses, pp. 161–176. Springer (2012)

12. Omohundro, S.M.: Autonomous technology and the greater human good. In: Müller, V.C. (ed.) Risks of artificial intelligence, pp. 9–27. CRC Press (2016)
13. Russell, S.J., Norvig, P.: Artificial intelligence. Prentice Hall (2010)
14. Tessier, C.: Robots autonomy. In: Lawless, W.F., Mittu, R., Sofge, D., Russell, S. (eds.) Autonomy and artificial intelligence, pp. 179–194. Springer International Publishing (2017)
15. Yampolskiy, R.V., Fox, J.: Artificial general intelligence and the human mental model. In: Eden, A.H., Moor, J.H., Søraker, J.H., Steinhart, E. (eds.) Singularity hypotheses, pp. 129–145. Springer (2012)
16. Yampolskiy, R.V., Fox, J.: Safety engineering for artificial general intelligence. Topoi **32**(2), 217–226 (2013)
17. Yudkowsky, E.: Creating friendly AI 1.0. The Singularity Institute (2001)
18. Yudkowsky, E.: Artificial intelligence as a positive and negative factor in global risk. In: Bostrom, N., Čirkovič, M.M. (eds.) Global catastrophic risks, pp. 308–345. Oxford University Press (2008)
19. Yudkowsky, E.: Complex value systems in friendly AI. In: Schmidhuber, J., Thrisson, K.R., Looks, M. (eds.) Artificial general intelligence, pp. 388–393. Springer (2011)
20. Yudkowsky, E.: Friendly artificial intelligence. In: Eden, A.H., Moor, J.H., Søraker, J.H., Steinhart, E. (eds.) Singularity hypotheses, pp. 181–193. Springer (2012)