Is there a trade-off between human autonomy and the ‘autonomy’ of AI systems?

Carina Prunkl
Institute for Ethics in AI, University of Oxford

Abstract

The development and deployment of artificial intelligence (AI) systems to perform a wide variety of tasks has raised new questions about how AI may affect human autonomy. Numerous guidelines on the responsible development of AI now emphasise the need for human autonomy to be protected. In some cases, this need is linked to the emergence of increasingly ‘autonomous’ AI systems that can perform tasks without human control or supervision. Do such ‘autonomous’ systems pose a risk to our own human autonomy? In this article, I address the question of a trade-off between human autonomy and system ‘autonomy’.

1 Introduction

Trading, driving, hiring, or firing—a growing number of tasks is now performed by automated decision-making software or robotic systems that no longer require the direct control or supervision of human operators. The increasing delegation of tasks to autonomous systems gives rise to a number of complex moral issues, including the question of how these developments affect our *own* abilities to make and execute decisions that are of practical import to our lives. In other words, it gives rise to concerns about human autonomy.

Human autonomy has become a central theme across guidelines and principles on the responsible development of artificial intelligence (AI). ‘Respect for autonomy’, for example, is the first of four key ethical principles of the European Commission’s High-Level Expert Group’s Ethics Guidelines for Trustworthy AI (High-level expert group on artificial intelligence, 2019, p.12) and it is the second principle of the Montreal Declaration for responsible development of artificial intelligence (Montreal, 2017). While the frequent call for the protection of human autonomy suggests consensus across guidelines, closer inspection reveals substantial heterogeneity as to what the perceived risks to human autonomy in fact are.

In some cases, the need for protection of human autonomy is linked to the increasing ability of AI systems to perform tasks 'autonomously', that is, independent of human control or supervision. Floridi & Cowls (2019), for example, write: "The risk is that the growth in *artificial autonomy* may undermine the flourishing of *human autonomy*" (original emphasis), or "[...] the autonomy of machines should be restricted and made intrinsically reversible, should human autonomy need to be protected or re-established [...]" (Floridi & Cowls, 2019, 3.3). Bayouth et al. (1997) even speak of a potential 'transfer' of human autonomy to autonomous systems.

This article examines the relationship between human autonomy and the 'autonomy' of autonomous systems (henceforth *system 'autonomy'*). In particular it addresses the question of whether we should expect an inherent *trade-off* between human autonomy and system 'autonomy'.

2 Two concepts: human autonomy and system 'autonomy'

Human autonomy

Human autonomy broadly refers to a person's effective capacity to act on the basis of beliefs, values, motivations, and reasons that are in some relevant sense her own (Christman, 2018; Mackenzie & Stoljar, 2000). Notably, this definition contains two distinct requirements for a person or for her actions to be autonomous (Prunkl, 2022). The first, *authenticity*, requires a person's beliefs and values to *actually be her own* and to not be the product of external manipulative or distorting influences. They need to be authentic, or in a relevant sense reflective of her 'inner self' (however one defines such an inner self) (Frankfurt, 1971; Dworkin, 1988; Christman, 2009). The second requirement, *agency*, is that a person has the *effective capacity* to enact decisions, make choices, and take charge of important aspects of their lives Mackenzie (2014). This implies that they must have both certain freedoms and meaningful options available to them that allow them to make choices about who to be and what to do (Mackenzie, 2014, p.17).

System 'autonomy'

In the context of AI research, the term 'autonomous' is used to describe systems that operate independently from or without the control of human operators (Franklin & Graesser, 1996). In technical circles, 'autonomy' is also sometimes associated with the ability of AI systems to learn and act on the basis of experience (Russell & Norvig, 1998; Wooldridge & Jennings, 1995). It is immediately evident that this notion of 'autonomy' only superficially (if at all) resembles what we take to be human autonomy. It is void of the complex cognitive processes that are necessary for the formation of beliefs and values, as well as consciousness and self-awareness, both of which are normally considered fundamental to autonomy (Commission et al., n.d.).[1] It also differs from human autonomy with respect to the role it plays within society. Human autonomy is often considered a fundamental value that grounds many of our moral and political institutions, at least within the context of Western society and politics (Christman, 2009). 'Autonomy', on the other hand, is at most of instrumental value to whatever task the AI system in question is designed to achieve.

3 Is there a trade-off?

To answer the question of whether there is a trade-off between human autonomy and system 'autonomy' it is helpful to consider in more detail how exactly AI systems might impact human autonomy. I distinguished above between authenticity requirements (i.e., the authenticity of beliefs) and agency requirements (i.e., the effective capacity to execute decisions of import). AI systems can impact the authenticity dimension of human autonomy in at least two ways: they can exert distorting influences on people, for example by subjecting them to personalised manipulative online content (Susser et al., 2019). On the other hand, AI systems can also have a positive impact on authenticity: they can help us think through complex decision-making processes and thus make better decisions, as well as help us to overcome potential biases, e.g. in the context of clinical decision-making. In both cases, it is not the 'autonomy' of an AI system that dictates whether or not human autonomy is undermined or strengthened, but the design of the system, the manner and context in which it is deployed. With respect to the authenticity dimension of autonomy, therefore, there is no *inherent* trade-off between system 'autonomy' and human autonomy.

More interesting is the question of how system 'autonomy' interacts with the *agency* dimension of human autonomy. When we delegate tasks to AI systems, do we give up autonomy by doing so? To get to the bottom of this question, it is useful to first consider a scenario that does not involve AI systems at all, but instead deals with humans and human interactions. Imagine Sarah asks her friend John to do the grocery shopping for her because she finds herself unusually busy. "Choose what you want", she tells him, and he goes ahead and buys either what he likes best or, ideally, what he thinks she might need or like. This is a clear case of task delegation: Sarah has delegated the task of grocery shopping to another person. By doing so, has she given up her autonomy? Hardly. In this example, nobody has forced Sarah to ask John to do the shopping, nor to eat what he bought. Delegation itself, therefore, is not tantamount to giving up autonomy, as long as the act is voluntary and

[1] This is not to claim that machines will never be able to obtain the capacity for autonomy.

based on informed consent.[2] This reasoning translates almost seamlessly into the context of AI: the mere delegation of tasks to *AI systems* is not tantamount to giving up autonomy.[3]

Yet, we often find ourselves in situations where 'voluntary' does not seem to adequately describe the choices at hand. What if Sarah really could not find the time to do the shopping and had no choice but to ask John for help? Does the necessity of the situation imply that Sarah was not acting autonomously? This raises questions about the role of *dependencies* in discussions about autonomy. Naturally, such dependencies can take on different forms: we might be dependent on artefacts, services, or other human beings. Dependence will also become increasingly important in the context of AI. Given the speed at which numerous tasks are being automated, it seems unavoidable that our overall dependence on AI driven systems will starkly increase. In itself, this is not necessarily problematic. We already are dependent on all sorts of things, people, and relationships. For example, we already depend on technologies such as fire or electricity. As social animals, we also depend on other humans in fundamental ways. And so if autonomy were to be equated with independence, hardly any human would ever qualify as autonomous.[4] What is relevant, instead, is whether these dependencies are legitimate and morally defensible.

Returning to the original issue of AI and the agency dimension of autonomy: it is necessary to understand how AI systems affect the freedoms and opportunities that we as a society consider fundamental. Whether a given AI system is acting 'autonomously' when it is impeding on these freedoms, e.g. through coercion, is secondary to the question of whether it's actions *are* impeding on them. In other words, there is no *inherent* trade-off between 'autonomy' and agency.

This is not to say that autonomous systems never pose a risk to autonomy. My argument only establishes that there is no *inherent* trade-off between human autonomy and system 'autonomy'. There are clearly cases in which autonomous systems can have a negative impact on human autonomy. Automated decisions of loan applications, for example, can seriously affect the options available to loan applicants, with potentially severe consequences for their life choices. Recommendation algorithms on social media platforms can manipulate users into spending more time on the platform than they would approve of, if they were given the opportunity for critical reflection. In both cases, the use of autonomous systems can be considered as autonomy-undermining. It is also conceivable that in particular risks

[2] The relationship between autonomy and informed consent is also frequently discussed in the context of biomedical ethics, e.g. (Beauchamp et al., 2001).

[3] In a similar vein, Floridi & Cowls (2019) emphasise the importance of humans to be able to freely choose which decisions are delegated to AI systems, as well as to be able to reverse this choice, if needed. The authors call this the 'decide-to-delegate' model. Notably, there exists some ambiguity regarding who these 'humans' are, that is, whether it refers to any or all humans, users, operators, citizens, etc. Depending on the answer, the 'deciding-to-delegate' model will result in radically different demands on system design or governance mechanisms.

[4] This has been emphasised many times in the literature on *relational autonomy*. See e.g. Hutchison et al. (2018) and references therein.

to the agency dimension of autonomy become more pressing with increasing deployment and development of AI systems. The argument made here, however, has not been that autonomous AI systems can never negatively affect human autonomy, but rather that they do not do so *in virtue* of their 'autonomy'. Risks to autonomy primarily emerge from how AI systems are used, the context in which they are deployed, and the unpredictable nature of some AI systems, rather than from the fact that these systems operate 'autonomously'.

References

Bayouth, M., Nourbakhsh, I., & Thorpe, C. (1997). A hybrid human-computer autonomous vehicle architecture. In *Proceedings, Third ECPD International Conference on Advanced Robotics, Intelligent Automation and Control.* Citeseer.

Beauchamp, T. L., Beauchamp, P. o. P. a. S. R. S. T. L., Childress, J. F., & Childress, U. P. a. H. P. o. E. J. F. (2001). *Principles of Biomedical Ethics.* Oxford University Press.

Christman, J. (2009). *The Politics of Persons: Individual Autonomy and Socio-Historical Selves.* Cambridge University Press.

Christman, J. (2018). Autonomy in Moral and Political Philosophy. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2018 ed.). Metaphysics Research Lab, Stanford University.

Commission, E., for Research, D.-G., Innovation, on Ethics in Science, E. G., & Technologies, N. (n.d.).

Dworkin, G. (1988). *The Theory and Practice of Autonomy.* Cambridge University Press.

Floridi, L., & Cowls, J. (2019, June). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, *1*(1).

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

Franklin, S., & Graesser, A. (1996). Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. In *International Workshop on Agent Theories, Architectures, and Languages* (pp. 21–35). Springer.

High-level expert group on artificial intelligence. (2019). *Ethics guidelines for trustworthy AI* (No. B-1049). Brussels.

Hutchison, K., Mackenzie, C., & Oshana, M. (2018). *Social Dimensions of Moral Responsibility.* Oxford University Press.

Mackenzie, C. (2014). *Three Dimensions of Autonomy: A Relational Analysis.* Oxford University Press.

Mackenzie, C., & Stoljar, N. (2000). *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self.* Oxford University Press.

Montreal. (2017). Montreal Declaration for Responsible Development of AI. *Forum on the Socially Responsible Development of AI*.

Prunkl, C. (2022, February). Human autonomy in the age of artificial intelligence. *Nature Machine Intelligence*, *4*(2), 99–101. Retrieved 2022-03-21, from `https://www.nature.com/articles/s42256-022-00449-9` doi: 10.1038/s42256-022-00449-9

Russell, S., & Norvig, P. (1998). *Artificial Intelligence: A Modern Approach* (2edition ed.). Upper Saddle River, NJ: Pearson.

Susser, D., Roessler, B., & Nissenbaum, H. (2019, June). *Technology, Autonomy, and Manipulation* (Tech. Rep.). Rochester, NY: Social Science Research Network.

Wooldridge, M., & Jennings, N. R. (1995). Agent theories, architectures, and languages: A survey. In M. J. Wooldridge & N. R. Jennings (Eds.), *Intelligent Agents* (pp. 1–39). Berlin, Heidelberg: Springer.