Human autonomy in the age of artificial intelligence

Carina Prunkl
Institute for Ethics in AI, University of Oxford

January 13, 2022

Abstract

Current AI policy recommendations differ on what the risks to human autonomy are. To systematically address risks to autonomy, we need to confront the complexity of the concept itself and adapt governance solutions accordingly.

It is hard to overstate the important role autonomy plays for our moral and political institutions. A cornerstone of human dignity and a prerequisite of liberal democracy, autonomy is often considered a fundamental human value [1–4]. Progress in artificial intelligence (AI) development opens up new opportunities for supporting and fostering autonomy, but it simultaneously poses significant risks. Recent incidents of AI-facilitated deception, manipulation, or coercion suggest that AI technologies could seriously interfere with human autonomy on a large scale. Cambridge Analytica's attempt to manipulate voters is just one example [5]. Facebook's "emotional contagion" experiment, in which users were swayed towards adopting certain emotional states, yet another one [6].

Consequently, human autonomy has become a central theme across guidelines and principles on the responsible development of AI. The European Commission's High-Level Expert Group (HLEG) lists 'respect for autonomy' as the first of its four key ethical principles in their Guidelines on Trustworthy AI [7]. Several other policy documents, including the Association for Computing Machinery's Code of Ethics [8], The Montreal Declaration for Responsible Development of Artificial Intelligence [9], and the European Commission's White Paper on Artificial Intelligence [10], equally emphasise the need to protect and respect autonomy whereas the Organisation for Economic Co-operation and Development (OECD) lists autonomy as one of their human-centered values [11].

Despite this frequent call for the protection of autonomy, there remains substantial ambiguity within these documents as to (a) what exactly is meant by the term 'autonomy', as well as (b) what the risks from AI to autonomy are. In some cases, 'autonomy' remains

undefined [8, 10]. Often, however, guidelines take different approaches to what they take the protection of human autonomy to entail. For example, the HLEG advocates that it entails no "unjustified coercion, deception, or manipulation" by AI systems [7], the OECD promotes "capacity for human determination" [11]. Others emphasise that "control over and knowledge about autonomous systems" [12] is needed and others again stress that principles of human autonomy translate into the protection of "human decision-making power" [13]. This is also consistent with findings by Fjeld et al., who found that autonomy typically provides the theoretical grounding for principles of 'human control of technology' [14].

The result of this heterogeneity is a patchwork of seemingly disjoint policy recommendations. To illustrate this point further: it is one thing to implement measures that protect users from fraudulent online manipulation (e.g. to prevent incidents like Cambridge Analytica), but it requires an entirely different set of measures to ensure human decision-making power (e.g. to ensure that the passenger of a driverless car has authority over most functions of the car). This poses a challenge to policy makers: how can we adequately address potential risks to human autonomy?

The overall lack of structure in the current discourse threatens to undermine ongoing governance efforts, efforts that are already straining under the complexity of the technical landscape and the large uncertainty of AI's social impacts. While there has been remarkable scholarly progress in individual areas, such as online manipulation [5, 15–18] or healthcare [19, 20], few scholars have discussed the concept of autonomy within a broader technological context [21–23]. To adequately address the risks AI might pose to human autonomy, we first need a clearer view of *what we mean* by 'human autonomy' and *how* AI technologies could interfere with it. The following aims to add structure to the debate by highlighting different dimensions of human autonomy, providing examples of how AI systems might interfere with them, and discussing some of the policy implications

Human autonomy as agency and authenticity

'Autonomy' is a notoriously complex concept [24, 25], but it generally can be taken to refer to a person's effective capacity for self-governance. This means that he or she can act on the basis of beliefs, values, motivations, and reasons that are in some relevant sense their own [3, 25]. There are (at least) two fundamental aspects to this definition, each pointing to a different set of conditions that need to be fulfilled for a person (or action) to count as autonomous:

1. **Authenticity**. The beliefs, values, motivations, and reasons held by a person are in a

relevant sense *authentic* to that person, i.e. not the product of external manipulative or distorting influences.

2. **Agency**. A person *is able* to act on the beliefs and values they hold. This implies that they have meaningful options available to them, allowing them to make choices that are of practical import to their life.

Distinguishing between authenticity and agency explains and clarifies some of the heterogeneity found in the current policy discourse. Those calling for protection from AI-facilitated manipulation and deception are primarily addressing the authenticity dimension of autonomy, whereas those emphasising the importance of retaining control over one's own decisions do so in reference to agency.

Here are some explicit examples for how AI systems could affect *authenticity*:

Manipulation is a form of external – often covert – influence by which people's decision-making vulnerabilities are targeted and exploited [26]. Through the analysis of large amounts of data, AI systems are able to identify such vulnerabilities and could be used to exploit them. Recommendation systems, often used by search engines and social media platforms, currently pose one of the highest risks for AI-facilitated online manipulation [5, 15–18].

Adaptive preference formation refers to the process of a person adapting their preferences so as to match the options that are available to them [27]. The increasing use of recommendation algorithms to pre-select online content or options can lead to such adapted preferences, as first studies suggest [28]. This phenomenon might be reinforced by automation bias, the tendency of humans to favour suggestions from computational systems.

Deception and adaptive belief formation are another way in which AI systems might affect authenticity, which relies on the availability of adequate information so as to make appropriate judgments. The amplification of conspiratorial content on social media platforms as a result of algorithmic content selection is an example for how AI systems participate in the shaping of beliefs.

Agency, on the other hand, might be negatively affected by the following:

Loss of opportunities. AI systems may create new opportunities for individuals to thrive, but they can also lead to a loss of opportunities, such as when automated decision-making algorithms are racially biased and prevent individuals from accessing health care [29].

Loss of freedom. AI might equally contribute to the restriction of basic liberties directly, e.g. through the deployment of military drones, or indirectly, e.g. through the enabling of large-scale surveillance.

Authenticity	Agency
Manipulation	Loss of opportunities
Adaptive preference formation	Loss of freedom
Deception and adaptive belief formation	Loss of competence
	Paternalism

Table 1: Selected risks to human autonomy.

Loss of competence to make decisions might occur if more and more tasks are routinely outsourced to AI systems, including decision-making in social, medical, or financial settings.

Paternalism involves well-intentioned infringements on a person's autonomy against their will [30]. AI systems that engage in full paternalistic behaviour are mostly future talk at this point but concerns about paternalism have already been raised in the context of health apps [31].

For a summary of the listed risks, see Table 1.

Policy challenges and implications

The above distinction between authenticity and agency can be re-captured by two main questions:

1. Does the use of a given AI system lead to the unwarranted distortion of an individual's beliefs, motivations, or decisions?
2. Does the use of a given AI system limit basic liberties or opportunities, or else prevents individuals from executing decisions of practical import to their lives?

Answering each question poses additional challenges to developers and policy makers: addressing authenticity requires prior deliberation on what conditions need to be fulfilled for external influence to count as (im)permissible. Addressing the external dimension, on the other hand, requires a decision on which options and freedoms are considered essential for autonomy. It also requires deliberation on the permissibly of potential trade-offs between such freedoms.

There exists an extensive body of philosophical literature that is concerned with the first challenge and explicitly lays out which conditions need to be fulfilled for a decision or desire to count as authentic. A prominent approach by Christman is to consider a person's decision or desire as authentic if and only if she does not feel *alienated* from the decision or desire, were she to critically reflect on them [32]. This account emphasises the importance of the individual's point of view when determining whether an external influence counts as

autonomy-undermining. Coming back to the context of AI, this points towards including *users* of AI systems much more into the discourse on human autonomy: to determine whether a given system (or the way it is used) is, say, manipulative, it does not suffice to merely observe user behaviour. Instead, we need to test whether users endorse their decisions when given the opportunity to critically reflect on them.

Addressing the second challenge will require explicitly laying out any freedoms, opportunities, or decisions that could be affected (positively or negatively; directly or indirectly) by the deployment of any given AI system. Trade-offs should be made explicit and citizens should be informed about any such limitations or trade-offs.

Identifying potential risks from AI development is a mammoth task. The uncertainty and complexity that surrounds ethical and social impacts of emerging technologies poses significant challenges to those involved in the governance process. Tackling these challenges requires us to be clear on what it is we are concerned about in the first place. Only then can we begin putting in place adequate governance mechanisms that prevent and mitigate potential negative impacts.

Acknowledgments

The author thanks Prof. John Tasioulas, Dr. Milo Philipps-Brown, Dr. Carissa Veliz, Dr. Ted Lechterman, Prof. Allan Dafoe, and Ben Garfinkel for their helpful comments. Funding: No external funding sources. Competing interests: The author declares that they have no competing interests.

Competing Interests

The author declares no competing interests.

References

[1] Joseph Raz. *The Morality of Freedom*. Clarendon Press, June 1986.

[2] Christine M. Korsgaard, Christine Marion Korsgaard, Korsgaard Korsgaard, Christine Marion, Gerald Allan Cohen, Raymond Geuss, Thomas Nagel, and Bernard Williams. *The Sources of Normativity*. Cambridge University Press, June 1996.

[3] John Christman. Autonomy in Moral and Political Philosophy. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2018 edition, 2018.

[4] Beate Roessler. *Autonomy: An Essay on the Life Well-Lived*. John Wiley & Sons, May 2021.

[5] Daniel Susser, Beate Roessler, and Helen Nissenbaum. Technology, Autonomy, and Manipulation. Technical report, Social Science Research Network, Rochester, NY, June 2019.

[6] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, June 2014. Publisher: National Academy of Sciences Section: Social Sciences.

[7] HLEG. Ethics guidelines for trustworthy AI. Technical Report B-1049, Brussels, 2019.

[8] Computing Machinery ACM. ACM code of ethics and professional conduct. *Code of Ethics*, 1992.

[9] Montreal. Montreal Declaration for Responsible Development of AI. *Forum on the Socially Responsible Development of AI*, 2017.

[10] ECWP. On Artificial Intelligence - A European approach to excellence and trust. White Paper COM(2020), February 2020.

[11] OECD. Recommendation of the Council on Artificial Intelligence. Technical Report OECD/LEGAL/0449, 2019. URL `https://oecd.ai/en/ai-principles`.

[12] ECAA. Statement on Artificial Intelligence, Robotics, and 'Autonomous' Systems. Technical report, 2018.

[13] Luciano Floridi and Josh Cowls. A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, 1(1), June 2019.

[14] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. SSRN Scholarly Paper ID 3518482, Social Science Research Network, Rochester, NY, January 2020. URL `https://papers.ssrn.com/abstract=3518482`.

[15] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. Recommender Systems and their Ethical Challenges. SSRN Scholarly Paper ID 3378581, Social Science Research Network, Rochester, NY, April 2019.

[16] Rafael A. Calvo, Dorian Peters, and Sidney D'Mello. When technologies manipulate our emotions. *Communications of the ACM*, 58(11):41–42, October 2015.

[17] Eliza Mik. The erosion of autonomy in online consumer transactions. *Law, Innovation and Technology*, 8(1):1–38, January 2016.

[18] Natali Helberger. Profiling and Targeting Consumers in the Internet of Things – A New Challenge for Consumer Law. Technical report, Social Science Research Network, Rochester, NY, February 2016.

[19] C. Burr, J. Morley, M. Taddeo, and L. Floridi. Digital Psychiatry: Risks and Opportunities for Public Health and Wellbeing. *IEEE Transactions on Technology and Society*, 1(1):21–33, March 2020.

[20] Jessica Morley and Luciano Floridi. The Limits of Empowerment: How to Reframe the Role of mHealth Tools in the Healthcare Ecosystem. *Science and Engineering Ethics*, 26(3):1159–1183, June 2020.

[21] Roger Brownsword. Autonomy, delegation, and responsibility: agents in autonomic computing environments. In *Law, Human Agency and Autonomic Computing*, pages 80–100. Routledge, 2011.

[22] Rafael Calvo, Dorian Peters, Karina Vergobbi Vold, and Richard Ryan. Supporting human autonomy in AI systems: A framework for ethical enquiry. Springer, 2020.

[23] Alan Rubel, Clinton Castro, and Adam Pham. *Algorithms and Autonomy: The Ethics of Automated Decision Systems*. Cambridge University Press, May 2021.

[24] Gerald Dworkin. *The Theory and Practice of Autonomy*. Cambridge University Press, August 1988.

[25] Catriona Mackenzie. *Three Dimensions of Autonomy: A Relational Analysis*. Oxford University Press, August 2014.

[26] Robert Noggle. Manipulative Actions: A Conceptual and Moral Analysis. *American Philosophical Quarterly*, 33(1):43–55, 1996.

[27] Jon Elster. *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge University Press, October 1985.

[28] Gediminas Adomavicius, Jesse C. Bockstedt, Shawn P. Curley, and Jingjing Zhang. Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects. *Information Systems Research*, 24(4):956–975, September 2013.

[29] Heidi Ledford. Millions of black people affected by racial bias in health-care algorithms. *Nature*, 574(7780):608–609, October 2019.

[30] Gerald Dworkin. Paternalism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2020 edition, 2020. URL `https://plato.stanford.edu/archives/fall2020/entries/paternalism/`.

[31] Michael Kühler. Exploring the phenomenon and ethical issues of AI paternalism in health apps. *Bioethics*, n/a(n/a), 2021.

[32] John Christman. *The Politics of Persons: Individual Autonomy and Socio-Historical Selves*. Cambridge University Press, 2009.