AI EXPLAINABILITY, TEMPORALITY, AND CIVIC VIRTUE

Wessel Reijers, Tobias Matzner, Suzana Alpsancar, Martina Philippi

Paderborn University (Germany)

Wessel.reijers@uni-paderborn.de; tobias.matzner@uni-paderborn.de; suzana.alpsancar@uni-paderborn.de; martina.philippi@uni-paderborn.de

EXTENDED ABSTRACT

The notion that artificial intelligence (AI) has to be explainable has become entrenched in the public discourse concerning the ethical impacts of this emerging technology (Mittelstadt et al., 2016). Most notably, the stated reason for this concern is the property of neural networks to function as 'black box' models (Pasquale, 2015) that nonetheless perform certain modalities of reasoning. That is to say, these models 'reason' from particular inputs, which may consist of characters, pixels, or digital information in other modalities, to particular outputs, without transparently disclosing the process of this reasoning. This is often contrasted with 'good old fashioned AI' (GOFAI) models that use decision trees which – in principle – can be followed by a human expert from input to output. The problem with neural nets, implemented in programs like ChatGPT and Dall-E, is that they can potentially influence or even autonomously make decisions about human affairs that cannot ex-post be explained by human interpreters – even if these are experts. At most, humans may figure out the particular artificial neurons that had an important influence on a decision.

Yet, the feasibility and relevance of the principle of explainability has been questioned. Robbins (2019) has argued that in fact, people are not required to explain every decision they make. Instead, explainability only becomes an issue in exceptional circumstances when the outcome of a particular decision requires explanation. It would therefore be unreasonable and unhelpful to insist on a standard for AI systems that does not apply to human decision-making. Moreover, meaningful human control over AI decision-making, which is arguably one of the aims of explainability, can be achieved by other means – for instance through proper legislation. Others have argued that explainability should not be reduced to explicability (i.e., accounting for the explanandum) but should involve the social context, considering it as a set of social practices (Rohlfing et al., 2021). Indeed, explaining takes place in a social context, and moreover has different modalities.

From this perspective, explainability as such is neither a mere technical matter, nor is it in any case relevant, nor is it a singular phenomenon. This paper proposes an initial way to grapple with these difficulties, by considering – first of all – the role of temporality in different modalities of explaining, and – secondly – the normative perspective of civic virtue to evaluate these different modalities, which then raises distinct requirements for explainability given distinct social contexts.

Let us start with the consideration of temporality, as it offers a ground to consider different modalities of explanation. In the *Rhetoric*, Aristotle set out the idea that argumentation occurs in different temporal modalities. It can be past-oriented, in which case it is *forensic*, explaining what has happened by reference to memory and traces. It can be present-oriented, in which case it is *epideictic*, explaining why a person or act deserves blame or honor, or the assignment

of virtue or vice. It can, furthermore, be future-oriented, in which case it is *deliberative*, explaining why particular future outcomes should or should not be supported. AI systems can, in principle, be involved in all three of these modalities of explaining, but they confront us with different normative requirements when they do. Forensic explanations, for instance, put forward requirements concerning historical proof, whereas deliberative explanations put forward requirements concerning (political) vision and conviction.

To make sense of these normative requirements, we may also draw from Aristotle. For in Aristotle, as Johnstone argues, (2023), ethics, rhetoric, and politics are fundamentally interrelated. Modalities of explanation, in other words, have a bearing on ethical and political life, in that they affect human virtues. Virtue is therefore a valid point of departure, as Vallor has forcefully argued (2016) in the context of technology ethics, in considering how AI affects explainability in a normative sense. Yet, virtue is also primarily grounded in the life of the individual, being anchored in *eudaimonia*, and does not yet offer the resources to bridge the gap between the ethics of the individual and the politics of the community. Civic virtue, developed in Aristotle's *Politics*, does offer this transitory concept, for it always mediates between the aim of the individual and the aim of the political community. As such, it is also inherently concerned with technology, as the technological infrastructure is a primary concern of the mode by which civic virtue is cultivated and enacted.

Strikingly, the distinct modalities of explanation and the distinct notions of civic virtue in political philosophy can each be grounded in a consideration of temporality. Like modalities of explanation, civic virtue can be past-, present-, and future-oriented. Past-oriented civic virtue finds its most vocal adherents in liberal and neo-republican thought, where it in an instrumental quality that draws from a history of reputational events, cultivating a sense of civility amongst a population (Pettit, 1997). Present-oriented civic virtue finds its footing in classical republican thought, where it requires institutional structures for the support of practices that aim at internal goods (MacIntyre, 2007). Future-oriented civic virtue finds its basis in existential republican thought, which puts forward the requirement of a durable public sphere that supports political action in concert (Arendt, 1958).

How do these different modalities of civic virtue help us to think through the modalities of explainable AI? First, they help us to consider the plurality of explanations insofar as they relate to different modalities of civic virtue. To give an example: when faced with a reputation-building AI (e.g., a credit scoring mechanism), the aim of such a system is to mediate past-oriented civic virtue; in that reputation building implies a historical record of reputational events. Such a mode of civic virtue put forward requirements deriving from forensic explanations. In other words, for such an AI to cultivate rather than to corrupt civic virtue, its explainability would need to safeguard requirements of – amongst others – historical proof. When faced with a more explicitly political AI (e.g., the use of AI in mass online deliberation), the aim of such a system is to mediate future-oriented civic virtue; in that it supports deliberative decision-making about alternative political pathways. Such a mode of civic virtue puts forward requirements deriving from deliberative explanations. Differently put, for such an AI to cultivate rather than to corrupt civic virtue, its explainability would need to respect requirements of – amongst others – political conviction. It goes without saying that the latter requirements would be rather more stringent and putting up a higher bar than the former.

What this tells us is, foremost, that not every explanation is equal. Whether an explanation is required at all, and what modality it should be in, depends on the temporal mode of the human

activities that an AI system affects. In a shorthand manner, one could argue that the more AI infringes onto the political realm, the more stringent explainability requirements will be. At the same time, the modality of those requirements will also change, for instance shifting from forensic to deliberative requirements.

KEYWORDS: Explainability, AI, civic virtue, temporality.

REFERENCES

Arendt, H. (1958). *The Human Condition* (Vol. 24, Issue 1). University of Chicago Press. https://doi.org/10.2307/2089589

Johnstone, C. L. (2023). *An Aristotelian Trilogy: Ethics, Rhetoric, Politics, and the Search for Moral Truth*.

MacIntyre, A. (2007). *After Virtue: A Study in Moral Theory*. University of Notre Dame Press.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data and Society*, *3*(2), 1–21. https://doi.org/10.1177/2053951716679679

Pasquale, F. (2015). *The Black Box Society*. Harvard University Press. https://doi.org/10.4159/harvard.9780674736061

Pettit, P. (1997). *Republicanism: A Theory of Freedom and Government*. Oxford University Press.

Robbins, S. (2019). A Misdirected Principle with a Catch: Explicability for AI. *Minds and Machines*, *29*(4), 495–514. https://doi.org/10.1007/s11023-019-09509-3

Rohlfing, K. J., Cimiano, P., Scharlau, I., Matzner, T., Buhl, H. M., Buschmeier, H., Esposito, E., Grimminger, A., Hammer, B., Hab-Umbach, R., Horwath, I., Hullermeier, E., Kern, F., Kopp, S., Thommes, K., Ngonga Ngomo, A.-C., Schulte, C., Wachsmuth, H., Wagner, P., & Wrede, B. (2021). Explanation as a Social Practice: Toward a Conceptual Framework for the Social Design of AI Systems. *IEEE Transactions on Cognitive and Developmental Systems*, *13*(3), 717–728. https://doi.org/10.1109/TCDS.2020.3044366

Vallor, S. (2016). *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford University Press.