https://doi.org/10.1038/s42256-022-00533-0

Distinguishing two features of accountability for AI technologies

Zoe Porter, Annette Zimmermann, Phillip Morgan, John McDermid, Tom Lawton and Ibrahim Habli

Policymakers and researchers consistently call for greater human accountability for AI technologies. We should be clear about two distinct features of accountability.

Across the AI ethics and global policy landscape, there is consensus that there should be human accountability for AI technologies[1]. These machines are used for high-stakes decision-making in complex domains – for example, in healthcare, criminal justice and transport – where they can cause or occasion serious harm. Some use deep machine learning models, which can make their outputs difficult to understand or contest. At the same time, when the datasets on which these models are trained reflect bias against specific demographic groups, the bias becomes encoded and causes disparate impacts[2–4]. Meanwhile, an increasing number of machines that embody AI, and specifically machine learning, such as highly automated vehicles, can execute decision-making functions and take actions independently of direct, real-time human control, in unpredictable conditions that call for adaptive performance. This development can make human agency seem obscure. Considering these problems, a heterogeneous group of researchers and organizations have called for stronger, more explicit regulation and guidelines to ensure accountability for AI and autonomous systems[1,5–7].

But what do we mean by 'accountability', and do we all mean the same thing? Accountability comes in different forms and varieties across rich and overlapping strands of academic literature in the humanities, law and social sciences. Scholars in the AI ethics field have recently proposed systematic conceptualizations of accountability to address this complexity[8–11]. Several researchers in the field[8,10] take explicit inspiration from Bovens's influential analysis of accountability as a social relation, in which he describes accountability as: "a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences"[12].

A welcome development within the AI ethics landscape would be greater conceptual clarity on the distinction between the 'explaining' and 'facing the consequences' features of accountability, as well as the relation between them.

This matters ethically, legally and politically, as these two core features of accountability – that is, giving an explanation, and facing the consequences – can come apart and pull in different directions. We highlight them because, as the quotation illustrates, they represent a central bifurcation of the concept of accountability[12,13]. In addition, their relation is particularly complex when it comes to AI technologies.

Two features of accountability

The first feature of accountability – the requirement to provide an explanation – is commonly highlighted and often prioritized in the AI ethics community[4,8,10,14]. Under its umbrella, wide-ranging research and policy interventions are being pursued, including explainable AI (XAI) techniques, dataset audits, audit trails, third-party audits, ethical 'black boxes', incident sharing databases and reporting obligations on system providers[6,7,15–18].

The second feature of accountability – facing the consequences – is also an urgent desideratum in the AI ethics field. Multidisciplinary agreement about the precise expression of this feature may not be straightforward. For example, among public accountability scholars and some moral philosophers, it is described as the possible imposition of sanctions[13,18,19]. But in the law, the notion of 'sanction' is so strongly tied to punishment and coercion[20] that this locution may have overly restrictive connotations – and this may also be how AI engineers approach the term.

We suggest one way forward would be simply to adopt a framing of the second feature of accountability in terms of actively being held responsible for outcomes. By this we mean being subjected to expressed responses and reactions from the forum. To draw an initial, central distinction, an actor might be held morally or legally responsible, depending on the social context in which they are operating and the standards against which they are judged. When held morally responsible, they might be blamed or reproached, lose the community's respect, face demands for apology or demands to make amends. But they might also be held morally responsible in positive ways, through rewards or public expressions of praise. When held legally responsible, an actor might be required to pay financial compensation, or be subject to a legal order, or face punishment. In what follows, we speak of such legal practices in terms of legal 'liability': being subjected to a legal power that has the potential to alter one's legal relations to other parties[21]. Mechanisms for holding people legally and morally responsible – fairly – have not yet sufficiently adapted to the deployment of AI technologies[22].

For ease of exposition, we call the first feature 'accountability (explanation)' and the second feature 'accountability(held responsible)'.

Mapping the relation between the two features of accountability

There is a relational core to accountability. One gives an explanation to people, for something. And one is held responsible by people, for something.

The two features of accountability are also inter-related. An explanation of how or why the design or use of a given AI technology led to objectionable outcomes is often necessary – albeit not sufficient – to establish who should be accountable (held responsible)[7,23,24]. Several papers presented at a leading conference in the subject, the ACM

Conference on Fairness, Accountability, and Transparency (FAccT), to varying degrees frame accountability(explanation) as a good that, among other things, will facilitate accountability(held responsible)[8–10,14,23,25–27]. But, as some researchers have started to describe[7,9,25], there is room for considerably more precision in the mapping between these two features of accountability. Here, we give three reasons why the project of achieving greater human accountability for AI calls for more clarity and precision in respect of both the distinction and the relation between these two features.

First, typically more actors will be accountable(explanation) than accountable(held responsible), given the variety of people from whom explanations are required; for example, after an accident. In the realm of legal responsibility, an actor might not owe the requisite legal duty to the victim to trigger liability, for instance, where the duty alleged to have been breached is instead held by their employer, but they may still have a duty to give evidence in court. And in cases where actions are distributed across actors, these two features of accountability may well fall on different individuals. Such types of dislocation between accountability(explanation) and accountability(held responsible) will also occur in the accountability ecosystem for AI.

Second, although accountability(explanation) is sometimes described as a desirable feature of AI technologies themselves[4,8], the machines do not possess the capacity to explain. Such descriptions are shorthand for the full description that there are certain techniques that enable people in wider socio-technical systems to explain the processes or outputs of the machines. This further demonstrates that the connection between accountability(explanation) and accountability(held responsible) is not a given but requires a proper ordering of accountability relationships.

Third, any assumption that there is a straightforward transition from accountability(explanation) to accountability(held responsible) is challenged by the observation that different kinds of accountability(explanation) inform different forms of accountability(held responsible) in different ways.

To make this clearer, let us distinguish between three kinds of account or explanation that may be given. Some accounts will provide descriptions of a decision, action or outcome (for example, 'the system recommended a treatment of insulin'). Some will give causal explanations, indicating what event produced a given effect (for example, 'the insulin caused the patient's glucose levels to go down'). Some accounts will furnish normative explanations, which are explanations of the reasons why a decision or action is right or good, or why it ought to be made or taken (for example, 'the prescription of insulin is right because simpler treatments have failed to control the disease and the use of insulin is likely to preserve life'). Accountability(explanation) needs in particular to provide comprehensive normative, reason-giving explanations or justifications.

XAI techniques are among the range of methods being developed under the aegis of accountability(explanation). They are not the only such method, but it is germane to the discussion to clarify which kinds of explanation they facilitate. XAI increases the ability of humans to provide descriptions of the model, such as which features of the input data (for example, modifiable factors such as the patient's current weight, or fixed ones such as their age) were most heavily weighted in the production of the output (such as the recommendation of insulin as a treatment). XAI can facilitate an approximation of the model's underlying logical or causal processes (for example, if x had not been the input, y would not have been the output)[28,29]. But XAI techniques do not in themselves contribute to a normative explanation, giving the reasons why the machine's outputs were the right ones or good ones in the circumstances (beyond the fact, for example, that most heavily weighted features were those that correlated with the 'desired' outputs when the system was being trained)[29]. These reasons must be given by the human developers of the model — or those who used the machine with a given purpose — within any guidelines set by regulation. One practical risk of failing to appreciate the limits of XAI in the provision of normative explanations is that people may be tempted to hide behind the data these techniques provide to avoid such questions altogether.

We earlier distinguished between two forms of accountability(held responsible): moral and legal. In general, when it comes to moral accountability(held responsible), although it is necessary to establish that the actor had some causal connection to an outcome, the main locus of scrutiny will be their normative explanations: the reasons they took to favour their conduct as right or good. If, for example, engineers can provide a truthful account that their decision to implement autonomous emergency braking in a car is made for good reasons, which affected parties could not reasonably reject, it would be unfair to blame them morally if their diligent implementation of this function leads to an injury in rare cases.

When it comes to legal accountability(held responsible), by contrast, normative explanations will only be required in some cases. In most cases, bare descriptions are required. Often, if not always, liability turns on a causal connection to the harm[30]. But cases of strict liability (that is, liability regardless of fault) do not call for the defendant to give normative explanations that their conduct was right or good and that they behaved as they ought to have done in the circumstances. All that is required in such cases is that they were appropriately related to the harm (for example, an employer's vicarious liability for their employee's torts (civil wrongs)). It is salient here to recognize, however, the justifications for strict liability encompass its pragmatic function, such as ensuring that victims receive compensation in practice[31].

These variations and complexities are important to clarify when we consider what kinds of accountability(explanation) are required to inform different forms of accountability(held responsible) for AI technologies.

The way forward

This is a concern of applied significance for real-world AI governance and regulation. Conceptual clarity around these two core and distinct features of accountability can sharpen practical reasoning. It can help to ensure that research and policy interventions are suitably targeted.

By contrast, failing to make the distinction between the two features could inadvertently provide a foil for actors to evade facing the consequences for adverse outcomes. Without explicit consideration of how account-giving can inform the practice of holding people responsible, actors may be encouraged to be transparent in ways that do not sufficiently open them up to the appropriate actions and responses of others[32]. Meanwhile, a prioritization of accountability(explanation) methods may in practice shift the spotlight off the fact that, for as long as it remains unclear how appropriately to hold actors responsible for harms caused or occasioned by AI, accountability(explanation) will not achieve the results that injured or wronged parties will seek.

As the cited recent scholarship shows, some excellent and detailed work on accountability is being done in the AI ethics field. To advance the conversation further, engineers, policymakers, lawyers, philosophers and social scientists not only need to be at the same table, but also need to be willing and able to construct a conceptually clear common language about accountability that readily facilitates deep and

productive multidisciplinary collaboration. The different ways in which 'facing the consequences' may be construed illustrates the challenge of this task. We have made one initial suggestion: to understand this second core feature of accountability in terms of being held morally and legally responsible by the forum.

Considerable progress would be demonstrated if it were possible to use this common language to articulate precisely the core parameters of interest in accountability for AI technologies and how these are related. Among other things, this will ensure greater clarity from ethicists and legal scholars on what is required to hold actors morally and legally responsible, and it will enable technical specialists and policymakers to show how the methods they develop or mandate meet these goals and needs.

Editor's note: this article has been peer reviewed.

Zoe Porter [1,2] ✉, Annette Zimmermann [3,4], Phillip Morgan [5], John McDermid[1,2], Tom Lawton [1,6] and Ibrahim Habli[1,2]

[1]Assuring Autonomy International Programme, Institute for Safe Autonomy, University of York, York, UK. [2]Department of Computer Science, University of York, York, UK. [3]Department of Philosophy, University of Wisconsin-Madison, Madison, WI, USA. [4]Carr Center for Human Rights Policy, Harvard University, Cambridge, MA, USA. [5]York Law School, University of York, York, UK. [6]Bradford Teaching Hospitals NHS Foundation Trust, Bradford, UK.

✉e-mail: zoe.porter@york.ac.uk

Published online: 22 September 2022

References

1. Fjeld, J. et al. *Berkman Klein Center Research Publication* No. 2020-1 (2020).
2. Barocas, S. & Selbst, A. D. *Calif. Law Rev* **104**, 671–732 (2016).
3. Zimmermann, A. & Lee-Stronach, C. *Can. J. Philos.* **52**, 6–25 (2021).
4. Kroll, J. A. et al. *Penn Law Rev* **165**, 633–705 (2017).
5. European Commission. https://www.aepd.es/sites/default/files/2019-12/ai-definition.pdf (2019).
6. *Algorithmic Accountability Act of 2022* 117th Congress, S.3572 (US Government, 2022).
7. Falco, G. et al. *Nat. Mach. Intell.* **3**, 566–571 (2021).
8. Kacianka, S. & Pretschner, A. In *Proc. 2021 ACM Conf. Fairness, Accountability, and Transparency* 424–437 (2021).
9. Cooper, A. F., Laufer, B., Moss, E. & Nissenbaum, H. In *Proc. 2022 ACM Conf. Fairness, Accountability, and Transparency* 864–876 (2022).
10. Wieringa, M. In *Proc. 2020 ACM Conf. on Fairness, Accountability, and Transparency* 1–18 (2020).
11. Donia, J. In *Proc. 2022 ACM Conf. on Fairness, Accountability, and Transparency* 598 (2022).
12. Bovens, M. *Eur. Law J.* **13**, 447–468 (2007).
13. Schedler, A. in *The Self-Restraining State: Power and Accountability in New Democracies* (eds Schedler, A. et al.) Ch. 2 (Lynne Rienner Publishers, 1999).
14. Kroll, J. A. In *Proc. 2021 ACM Conf. on Fairness, Accountability, and Transparency* 758–771 (2021).
15. Falco, G. & Siegel, J. *SAE Int. J. Transp. Cyber. & Privacy* **3**, 97–111 (2020).
16. Winfield, A. & Jirotka, M. in *Towards Autonomous Robotic Systems* (eds Gao, Y. et al.) 10454 (Springer, 2017).
17. Avin, S. et al. *Science* **374**, 1327–1329 (2021).
18. Mansbridge, J. in *The Oxford Handbook of Public Accountability* (ed. Bovens, M.) Ch. 4 (Oxford Univ. Press, 2014).
19. Watson, G. *Philos. Top.* **24**, 227–248 (1996).
20. Law, J. (ed.). *Oxford Dictionary of Law* (Oxford Univ. Press, 2022).
21. Hohfeld, W. N. *Yale Law J.* **26**, 710–770 (1917).
22. Artificial Intelligence Select Committee. *AI in the UK: ready, willing and able?* (UK House of Lords, 2018).
23. Raji, I. D. et al. In *Proc. 2020 ACM Conf. on Fairness, Accountability, and Transparency* 33–44 (2020).
24. Yeung, K. in Yeung, K. & Lodge, M. *Algorithmic Regulation* Ch. 2 (Oxford Univ. Press, 2019)
25. Fraser, S., Simcock, R. & Snoswell, A. In *Proc. 2022 ACM Conf. on Fairness, Accountability, and Transparency* 185–196 (2022).
26. Irion, K. In *Proc. 2022 ACM Conf. on Fairness, Accountability, and Transparency* 1561–1570 (2022).
27. Lima, G., Grgić-Hlača, N., Jeong, J. K. & Cha, M. In *Proc. 2022 ACM Conf. on Fairness, Accountability, and Transparency* 2013–2113 (2022).
28. Wachter, S., Mittelstadt, B. & Russell, C. *Harv. J. Law Technol.* **31**, 841–888 (2017).
29. McDermid, J. A., Jia, Y., Porter, Z. & Habli, I. *Philos. Trans. R. Soc. A* **379**, 20200363 (2021).
30. Hart, H. L. A. & Honoré, T. *Causation in the Law* (Oxford Univ. Press, 1985).
31. Cappelletti, M. *Justifying Strict Liability: a Comparative Analysis in Legal Reasoning* (Oxford Univ. Press, 2022).
32. Ananny, M. & Crawford, K. *New Media Soc.* **20**, 973–989 (2018).

Acknowledgements

We thank P. Noordhof and T. Stoneham for their comments. This work was supported by the Engineering and Physical Sciences Research Council (EP/W011239/1) and the Assuring Autonomy International Programme, a partnership between Lloyd's Register Foundation and the University of York.

Competing interests

T.L. is Head of Clinical AI at Bradford Teaching Hospitals NHS Foundation Trust. The remaining authors declare no competing interests.

Additional information

Peer review information *Nature Machine Intelligence* thanks Jacob Metcalf and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.