Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/CLSR

Computer Law & Security Review

Comment

Exclusivity and paternalism in the public governance of explainable AI ☆

Perry Keller [a,*], *Archie Drake* [b]

[a] *School of Law, King's College London, United Kingdom*
[b] *Research Associate, School of Law, King's College London*

ARTICLE INFO

Keywords:
Artificial intelligence
Explainability
Trust
Governance

ABSTRACT

In this comment, we address the apparent exclusivity and paternalism of goal and standard setting for explainable AI and its implications for the public governance of AI. We argue that the widening use of AI decision-making, including the development of autonomous systems, not only poses widely-discussed risks for human autonomy in itself, but is also the subject of a standard-setting process that is remarkably closed to effective public contestation. The implications of this turn in governance for democratic decision-making in Britain have also yet to be fully appreciated. As the governance of AI gathers pace, one of the major tasks will be ensure not only that AI systems are technically 'explainable' but that, in a fuller sense, the relevant standards and rules are contestable and that governing institutions and processes are open to democratic contestability.

In this comment, we address the apparent exclusivity and paternalism of goal and standard setting for explainable AI and its implications for the public governance of AI. We argue that the widening use of AI decision-making, including the development of autonomous systems, not only poses widely-discussed risks for human autonomy in itself, but is also the subject of a standard-setting process that is remarkably closed to effective public contestation. The implications of this turn in governance for democratic decision-making in Britain have also yet to be fully appreciated. As the governance of AI gathers pace, one of the major tasks will be ensure not only that AI systems are technically 'explainable' but that, in a fuller sense, the relevant standards and rules are contestable and that governing institutions and processes are open to democratic contestability.

In other work, we have asserted – and continue to investigate – the idea that UK AI governance is paternalist in nature.[1] This assertion might seem surprising in the face of the current vigorous debate over the importance of AI explainability in multiple domains. Explainable AI is undoubtedly a rational solution to the confidentiality, complexity and opacity problems that render wide public access to and understanding of AI decision-making impractical, if not impossible. Building reliable explainability into the functioning of AI systems will certainly improve the possibilities for autonomy in per-

[1] Perry Keller, 'Participatory Accountability at the Dawn of Artificial Intelligence', King's College London Law School Research Paper No. 2019-31 https://ssrn.com/abstract=3448315.

☆ Supported by EPSRC Grant EP/R033722/1 (Trust in Human-Machine Partnership).
* Corresponding author: Perry Keller, School of Law, King's College London, United Kingdom.
E-mail address: perry.keller@kcl.ac.uk (P. Keller).

https://doi.org/10.1016/j.clsr.2020.105490
0267-3649/

sonal decision making, especially where AI has a relatively high impact on peoples' lives.[2] In the best of outcomes, such 'human-centred' explainability will foster trust and genuine trustworthiness, which will promote the public legitimacy of AI decision-making.[3] Achieving that virtuous circle is the challenge of the moment.

It is worth considering how challenging this aspiration is proving in practice. Explainable AI must stretch the technical and commercial constraints on creating workable AI systems, to meet the developing principles and rules that will govern AI conduct. Those constraints are certainly more formidable than public debate sometimes acknowledges. Creating workable forms of explainability is not just technically challenging, but even impossible for some forms of AI.[4] In highly competitive market conditions, tech firms are moreover wary of disclosing trade secrets or other confidential information through explainability as well as the increasing costs of AI regulation.[5] How confident are we that the messy, obscure trade-offs that are likely to result will actually tend to generate trust and legitimacy?

1. Public governance of explainable AI: coherence and exclusivity

On the public governance side, explainability as a solution to the 'black box' problem is being quickly absorbed into legal and ethical thinking.[6] In both spheres, the potential harms of AI applications engage complex questions of fundamental values and rights. In information law, data protection has provided a key framework for subjecting automated decision-making to specific rights and duties, which are directly rooted in fundamental rights.[7] Other legal fields, from contract to competition law, are also widening in scope to address need to balance AI's potential benefits against its risks of harm.[8] The eruption of AI as a major public policy issue has also fuelled a global proliferation of AI ethical guidelines.[9] Indeed, Charles Raab asserts that '[t]here has been a noticeable 'turn' from reliance on legal regulation to an emphasis on ethics – and accountability and transparency as well – in this part of the field of information policy'.[10] Explainable AI as a public governance question has consequently become an increasingly congested legal and ethical challenge.

Consequently, standard-setting for explainable AI has a remarkably high coherence ambition, which aims for 'end to end' explainability. That is to say, explainability must be simultaneously suited to the needs of AI developers, users and human subjects, while also being simultaneously coherent technically, commercially, legally and ethically. Coherence in this sense means that all principles, rights and duties are sufficiently factored into governance's high demands on AI in a manner that is also technically and commercially practicable. Put this in perspective, RegTech and other forms of techno-regulation, in which regulator and regulatee AI systems are intermeshed, depend on broad and deep standards coherence.[11] In a future of ubiquitous AI autonomous systems, explainability will need to function coherently throughout various stages and levels of interaction with and around AI systems.[12]

The high coherence ambition of explainable AI demands a workable reconciliation between explainability's technical and commercial limitations and an array of public governance standards. The latter cannot overwhelm the former, but the former must be seen to abide by the latter. Plainly, that reconciliation will not be sustainable unless the boundaries for the substantive demands of public governance on explainability are clarified. There are obvious pressures, for example, to avoid defining personal autonomy and dignity needs of human subjects of AI decision-making in ways that disproportionately obstruct the basic viability of AI systems.[13] The much-discussed General Data Protection Regulation (GDPR) Article 22 'right to explanation' is thus limited in scope to ex-

[2] ICO and Alan Turing Institute, 'Explaining decisions made with AI', May 2020, https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-artificial-intelligence/.

[3] European Commission, High-Level Expert Group on Artificial Intelligence, 'The Ethics Guidelines for Trustworthy Artificial Intelligence', April 2019, https://ec.europa.eu/futurium/en/ai-alliance-consultation; Upol Ehsan and Mark O. Riedl, 'Human-centered Explainable AI: Towards a Reflective Sociotechnical Approach', arXiv:2002.01092 [cs.HC], February 2020.

[4] Hamon, R., Junklewitz, H. and Sanchez Martin, J., 'Robustness and Explainability of Artificial Intelligence, European Union, Luxembourg, 2020, https://ec.europa.eu/jrc/en/publication/robustness-and-explainability-artificial-intelligence.

[5] Brkan and Bonnet, 'Legal and Technical Feasibility of the GDPR's Quest for Explanation of Algorithmic Decisions: of Black Boxes, White Boxes and Fata Morganas', (2020) 11 *European Journal of Risk regulation* (1).

[6] Roger Brownsword, Eloise Scotford and Karen Yeung (eds), *The Oxford Handbook of Law, Regulation and Technology* (Oxford: Oxford University Press, 2017).

[7] Margot Kaminski, 'The Right to Explanation, Explained', (2019) 34 *Berkeley Technology Law Journal*, 1; Sandra Wachter and Brent Mittelstadt, 'A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI', (*Columbia Business Law Review,* 2019).

[8] European Commission, Expert Group on Liability and New Technologies, 'Report on Liability for Artificial Intelligence and other emerging digital technologies' (2019) https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeetingDoc&docid=36608; Centre for Data Ethics and Innovation, AI Barometer Report, June 2020, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/894170/CDEI_AI_Barometer.pdf.

[9] Jessica Fjeld and others, 'Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI', 2020.

[10] Charles D. Raab, *Information Privacy: Ethics and Accountability* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3057469.

[11] Eva Micheler and Anna Whaley, 'Regulatory Technology: Replacing Law with Computer Code', (2020) 21 *European Business Organization Law Review*, 349–377.

[12] Burton, Simon; Habli, Ibrahim; Lawton, Tom; McDermid, John Alexander; Morgan, Phillip David James; Porter, Zoe Larissa Mayne, Mind the Gaps: Assuring the Safety of Autonomous Systems from an Engineering, Ethical, and Legal Perspective, (2019) *Artificial Intelligence,* 279.

[13] Anna Jobin, Marcello Ienca, and Effy Vayena, 'The Global Landscape of AI Ethics Guidelines', *Nature Machine Intelligence*, 1.9 (2019), 389–99.

planations necessary for the exercise of the rights and remedies available to data subjects, while also potentially subject to a range of permitted exceptions to transparency.[14] It is not a public right to be given a fully comprehensive or systemic explanation of how an AI system generated a particular decision.

On the face of it, this is unexceptional. In standard setting for new technologies, the state is expected to dominate and, moreover, effective governance of AI technologies is likely to require significant exclusivity and paternalism.[15] Given explicit economic goals, it is unsurprising that the government may favour collaborations with firms over those with civil institutions. Key regulators, empowered and limited by legislation, will give principles and standards meaning in practice.[16] There have of course been parliamentary inquiries and consultation exercises. Drawing on societal values expressed in fundamental rights, the courts can also be expected to join in shaping the demands of public governance on AI explainability.

On the other hand, the extent of this exclusivity and paternalism in the development of UK AI governance is also genuinely remarkable. It runs counter to trends across liberal democracies towards widening the avenues for active public participation in policymaking, not least through freedom of information rights and innovations in judicial review. The direction of travel seems to undercut the prospects for the virtuous circle of human-centredness and trust mentioned above; without broader participation, it is unclear how an 'authorizing environment' of legitimacy and support will be achieved.[17]

2. Governance without contestability and contestation

The reasons for this shift in governance back towards historic expectations of exclusivity, paternalism and even deference are twofold. The first is a consequence of the societal shift towards reliance on complex, interconnected technologies in every aspect of human life. In these circumstances, direct public participation in standard-setting is impractical and burdensome. The complexity and opacity of AI systems, which is often daunting for AI specialists, is well beyond the comprehension of ordinary members of the public. The economic and security consequences of disclosing confidential information are, moreover, seemingly too high to permit anything but controlled public consultation. We will address these pragmatic objections in our conclusions.

The second reason concerns the impact of AI's complexity and opacity on the effectiveness of public information access rights and, in particular, the importance of rights to explainability. As noted above, a major purpose of AI explainability is to enhance the trustworthiness and legitimacy of AI systems by rendering at least some AI decision-making sufficiently understandable to stakeholders.[18] One key question is therefore who should be empowered to require that a particular AI application be rendered explainable. This is undoubtedly a power necessary for effective regulatory supervision and control of AI systems, for example including the work of the Financial Conduct Authority and the Information Commissioner's Office.[19] On the other hand, a regulator's power to compel explanations typically comes with significant safeguards for any disclosure of confidential information as well as duties to temper regulatory oversight to suit levels of risk.[20]

Public rights to explanation applicable to automated decision-making are, in contrast, highly unusual. As mentioned above, Article 22 of the GDPR confers only a limited right. That said, it does offer the possibility of rendering some AI-driven decision-making modestly transparent and even potentially accountable to individuals who are significantly harmed. In contrast, the other transparency rights and duties of the GDPR only concern 'personal data', which is existing information relating to a data subject. Plainly, a right to reasonably available, existing information will often be inadequate when seeking to understand the reasons why an AI system has produced a particular decision. What is needed is a right to compel an explanation.

In terms of public governance, the difference between rights to information and rights to explanation are of historic importance. Direct public rights to access information and to compel explanations first emerged in Victorian reforms to the disclosure rules of civil litigation and, much later, for disclosure in judicial review. While these litigant rights can potentially be used to force the disclosure of evidence necessary to advance specific litigation, they are subject to strict confidentiality and collateral use restrictions. Save for evidence subsequently disclosed to the public through court proceedings, information disclosed to other parties cannot normally be used to inform the public. The point here is that, while litigation disclosure rules have the potential to compel considerable AI explainability in the future, litigation only provides a narrow, albeit powerful, avenue into matters of public concern.

It was only through the Freedom of Information Act (FOIA) rights that the public gained an unsupervised right to compel

[14] L. Edwards and M. Veale, 'Slave to the Algorithm? Why a 'right to an Explanation' is Probably not the Remedy you are Looking for', (2017) 16 Duke Law and Technology Law Review, 17.

[15] Shirley Pearce, 'AI in the UK: The Story So Far', Committee on Standards in Public Life Blog, 19 March 2020, https://cspl.blog.gov.uk/2020/03/19/ai-in-the-uk-the-story-so-far/.

[16] See, for example, Bank of England and Financial Conduct Authority, 'Machine learning in UK financial services', October 2019 https://www.fca.org.uk/publication/research/research-note-on-machine-learning-in-uk-financial-services.pdf.

[17] Mark H. Moore (2013), Recognizing Public Value, Harvard University Press.

[18] European Commission, High-Level Expert Group on Artificial Intelligence, 'The Ethics Guidelines for Trustworthy Artificial Intelligence', April 2019, https://ec.europa.eu/futurium/en/ai-alliance-consultation.

[19] *See, for example, the powers of the Financial Conduct Authority to compel a person subject to investigation to attend and answer questions under the Financial Services and Markets Act 2000, 2000 C.8, s 171; See also,* the expanded powers of the Information Commissioner's Office created under Part 5 and Schedules 12-15 of DPA 2018.

[20] On the lifetime confidentiality obligations of United Kingdom government employees, see, Civil Service code, published as statutory guidance under S.5. Constitutional Reform and Governance Act 2010, 2010 c. 25;.

the disclosure of information held by public authorities.[21] The FOI access right considerably enlarged the scope for individuals or private entities to drive transparency in governmental decision making and also widened the possibilities for radically shifting the agenda in public affairs.[22] The Freedom of Information Act is, of course, a work of carefully constructed compromises. To minimise the risks of damaging disclosure of confidential information or overwhelming central and local government with impractically burdensome requests, the Act not only brims with overlapping exemptions, but also strictly limits the scope of the FOI access right. It is simply a right to existing information, entailing no duty to create information and no duty to explain.

Despite these structural compromises, FOIA changed the character of public governance in the United Kingdom. While a public authority could not be compelled to explain its decisions, FOIA could be used to force the disclosure of the information that was used to make the decision. The rationality of outcomes could at least be assessed by evaluating the factors taken into account in the decision-making process. In opening this potential route into the heart of governmental decision-making, the FOIA information access right has unmistakeable links with ideas of deliberative and participatory democracy.[23] As a right limited to existing information, the FOIA right will often fail to break through the complexity and opacity of AI decision-making. Nonetheless, it does operationalise and demonstrate the value of the idea that decision-making of public importance should be contestable and open to recurring public contestation.[24]

3. Conclusion

Information law, which concerns access, control and use of information, is being re-made through the impact of AI applications.[25] AI's confidentiality, complexity and opacity characteristics are becoming an accepted barrier to direct public enquiry, defeating the contestability that democratic government requires. Paternalist concern by legislators and regulators is, however, not an adequate substitute for engaged citizens who wish to advance dissenting views and challenge the definitions of AI risk and harm imposed upon them. More particularly, in striving to achieve the high coherence demands of explainable AI, legislators and regulators are unlikely to answer fully the questions of explainable AI for what purposes and explainable AI for whom.

The path towards less exclusive standard-setting for explainable AI is undoubtedly fraught with difficulties. The pragmatic objections to opening avenues for direct public participation in AI governance, discussed above, must be taken seriously. A new FOIA style 'right to explanation', for example, would unleash potentially overwhelming compliance burdens and confidential information disclosure risks. Nonetheless, it is critically important to challenge the convenient idea that the complexity and opacity of data analytics precludes public participation in AI governance. A more promising avenue may lie in the intermediary ground between regulators and public interest organisations, or individuals, technically qualified to ask the hard questions. Extending the idea of sandboxes and other regulatory spaces in which policies and decisions can be subject to third party contestation as part of the regulatory process may provide a new avenues for democratic participation in AI governance.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

[21] Freedom of Information Act 2000 (FOIA) C.36, s 1 – General right of access to information held by public authorities.

[22] B. Worthy and R. Hazell, 'Disruptive, Dynamic and Democratic? Ten Years of FOI in the UK', *Parliamentary Affairs*, Volume 70, Issue 1, 1 January 2017, 22, 40; B. Worthy, 'Freedom of Information and the Media', 60 (H. Tumber and S. Waisbord ,eds), *The Routledge Companion to Media and Human Rights*, (Routledge, 2017) 60; M. Schudson, *The Rise of the Right to Know: Politics and the Culture of Transparency,* 1945–1975, (Belknap Press, 2015).

[23] Stephen Elstub, 'Deliberative and Participatory Democracy' in (Andre Bächtiger, John S. Dryzek, Jane Mansbridge, and Mark Warren, ed's) The Oxford Handbook of Deliberative Democracy, 2018.

[24] Deirdre K. Mulligan, Daniel Kluttz, and Nitin Kohli, 'Shaping Our Tools: Contestability as a Means to Promote Responsible Algorithmic Decision Making in the Professions, (2019) https://ssrn.com/abstract=3311894.

[25] Julie E. Cohen, *Between Truth and Power: The Legal Constructions of Informational Capitalism* (Oxford University Press, 2019), Chapter One.