The European Journal of International Law Vol. 29 no. 4

Accountability for Governance Choices in Artificial Intelligence: Afterword to Eyal Benvenisti's Foreword

Lorna McGregor*

Abstract

A growing body of literature examines how to make the use of new and emerging technologies more transparent and explainable as a means to ensure accountability for harm to human rights. While a critical part of accountability, a predominant focus on the technology can result in the design and adaptation of accountability principles to 'manage' the technology instead of starting from an assessment of the governance choices actors make when integrating new and emerging technologies into their mandates. Recognition of the governance choices underpinning the introduction of new and emerging technologies is often overlooked in scholarship and practice. Yet, without explicit recognition of the role played by technology in governance, the disruptive effects of technology on (global) governance may be underplayed or even ignored. In this response, I argue that if the 'culture of accountability' is to adapt to the challenges posed by new and emerging technologies, the focus cannot only be technology-led. It must also be interrogative of the governance choices that are made within organizations, particularly those vested with public functions at the international and national level.

1. Introduction

In a rich Foreword to this volume, Eyal Benvenisti traces the evolution of a 'culture of accountability' in global governance, particularly of international organizations. He expresses concern that these tools, while still evolving, may become 'redundant' in a

* Lorna McGregor, Professor of International Human Rights Law and Director, Human Rights Centre, University of Essex. Principal Investigator, Human Rights, Big Data and Technology Project. Email: lmcgreg@essex.ac.uk.
This work was supported by the Economic and Social Research Council [grant number ES/M010236/1]. Particular thanks to Dr Daragh Murray and Vivian Ng for their insightful comments on this article.

context in which new and emerging technologies play a central role in governance structures.[1] He observes that this is both because of the power of global technology companies and the nature of these technologies.[2] His Foreword raises the critical question of whether accountability principles, such as those embodied by global administrative law, but also embedded in international human rights law, and the rule of law more generally, can effectively adjust and adapt to this new context.[3]

In this response, I suggest that the way in which the employment of new and emerging technologies is understood and framed is central to the sustainability and adaptability of accountability principles. The effects these technologies can have on the rights of individuals and groups are increasingly acknowledged. Initially, the focus centred on the risks to privacy but has now expanded to recognize the potential for discrimination and inequality as well as the wider threats to all human rights posed by new and emerging technologies.[4] A growing body of literature critically examines the possibilities for addressing these risks through the lens of the technology itself. For example, the literature on 'algorithmic accountability' questions whether and how algorithms can be made more transparent and explainable in order to facilitate accountability when their use adversely affects human rights or causes other types of societal harms.[5]

Addressing the constraints of the technology constitutes a critical component to building an effective accountability framework. However, if the sole focus, the risk arises that the governance choices actors – particularly those with public functions – make, in integrating new and emerging technologies into their mandates are overlooked, and therefore not subject to a critical and accountable lens. Without recognizing the role of technology in governance, as Benvenisti documents, the disruptive effects of technology on (global) governance may be underplayed or even bypassed. This has a direct effect on the accountability framework, which may then result in the design and adaptation of accountability principles to 'manage' the technology rather than starting from an assessment of the governance choices enabled by the new technology. Recognition of the governance choices underpinning the introduction of new and emerging technologies is often overlooked in scholarship and practice. However, in this response, I suggest that if the 'culture of accountability' is to adapt to the challenges posed by new and emerging technologies, the focus cannot only be technology-led. It must also be interrogative of the choices that are made within the governance of organizations, particularly those vested with public functions at the international and national level.

1 Benvenisti, 'Upholding Democracy Amid the Challenges of New Technology: What Role for the Law of Global Governance?' 29 *European Journal of International Law (EJIL)* (2018) 9 at 55, at Section 4C.

2 *Ibid.*, at 66.

3 See Harlow, 'Global Administrative Law: The Quest for Principles and Values' 17 *EJIL* (2006) 187.

4 L. McGregor, 'Cambridge Analytica is more than a data breach – it's a human rights problem' *The Conversation* (4 June 2018), available at http://theconversation.com/cambridge-analytica-is-more-than-a-data-breach-its-a-human-rights-problem-96601.

5 See the literature *infra* notes 6 and 7.

2. Technology-Led Accountability as an Incomplete Approach

A burgeoning literature addresses the effects that using algorithms in decision-making can have on those subject to the decision.[6] The literature focuses on ways in which to make algorithms more transparent and 'explainable' as a means to aid accountability for harm caused by their use.[7] Benvenisti builds on this literature in two key ways. First, he highlights the role of human dignity in any decision-making process, and notes the challenges posed in this regard by big data-driven algorithms.[8] Algorithms work on the basis of correlation not causation and produce outputs at a group or population level, but which are not determinative in relation to specific individuals. Benvenisti points out that such categorization of 'individuals into groups based on predetermined factors – in other words, based on the stereotyped objectifying of human beings' is '[d]irectly at odds with the very notion of human dignity – the understanding that the law must treat each individual as being unique'.[9]

Second, he highlights the impact of the use of algorithms within decision-making on the role of discretion, which he characterizes as a central feature of global administrative law. He argues that the use of algorithms in decision-making distorts the duty to exercise discretion with an open mind.[10] Given the nature of algorithms and the manner in which they work – which, as indicated above, operates on the basis of population-level not individual-specific analysis – there is a risk that they fail to take into account the unique characteristics of a specific individual which is a key aspect of discretion within decision-making.

Most of the literature on 'algorithmic accountability' acknowledges the risks and challenges posed by such technology and asks how they might be addressed through adaptive techniques. This includes analysis of ways to preserve human dignity and discretion within decision-making processes that are made or supported by algorithms. This is a critical aspect of the accountability framework but one which is built around the acceptance of the involvement of technology and adaptation of accountability principles to it. It therefore does not start from the question of whether actors are making a governance choice about the removal, reduction or reconfiguration of discretion

6 Mittelstadt *et al.*, 'The Ethics of Algorithms: Mapping the Debate', 3(2) *Big Data and Society* (2016); D. Kehl, *et al.*, 'Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing', *Responsive Communities* (2017); Kroll *et al.*, 'Accountable Algorithms', 165(3) *University of Pennsylvania Law Review* (2017) 633.

7 Kroll *ibid.*; Ananny and Crawford, 'Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability', 20 *New Media and Society* (2016) 973; Citron and Pasquale, 'The Scored Society: Due Process for Automated Predictions', 89(1) *Washington Law Review* (2014) 1; Zarsky, 'The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making', 41(1) *Science, Technology and Human Values* (2016) 118; Diakopoulos, 'Algorithmic Accountablity: Journalistic Investigation of Computational Power Structures, 3(3) *Digital Journalism* (2015) 398.

8 Benvenisti, *supra* note 1, at 54.

9 *Ibid.*, at 54.

10 *Ibid.*, at 55.

and whether this choice should be subject to a process of scrutiny and examination. In any other context, a departure (or even quasi-departure) from established ways of working – particularly by a public body – would require public scrutiny and justification through an accountability process. This is because methodologies, like discretion, embody key standards of fairness and recognized modes of governance within society. Yet, where the departure is manifested through technology, the shift may be minimized and go unrecognized due to the tendency to treat technology instrumentally.

Technology-led approaches to accountability are particularly evident in debates on humans 'in' and 'on' the loop. For example, arguments have been made that discretion is not affected – or the effects are mitigated – where a human is still either 'in' or 'on the loop'.[11] A human is 'in the loop' where the human is the actual decision-maker, and their decision is merely informed by the algorithm. In principle, the human retains the ability to introduce discretion regardless of the conclusion reached by the algorithm as a piece of supporting evidence. Concerns arise, however, that the human decision-maker may defer to the conclusion reached by the algorithm or afford it significant weight due to the purported scientific calculations it makes.[12] The risk of deference is likely to increase in relation to 'higher stakes' decisions. For example, in the context of a parole decision for a person convicted of a serious crime, a human may be reluctant to overturn a high algorithmically produced risk score, given the potential consequences that the person may re-offend. In such a situation, the decision of the human would likely be scrutinized, including a requirement to explain why they 'went against' the findings of the algorithm.

A human 'on the loop' is where the algorithm is the actual decision-maker but a human reviews its decisions. In principle, discretion could be applied by the human reviewer to challenge the algorithm's findings. However, this would first depend on the terms of the review which may only be aimed at identifying cases of clear unfairness or discrimination, for example, rather than a wholesale review of the facts. Where the reviewer is not able to review the facts and evidence afresh, the likelihood of discretion regularly playing a part in the decision is remote.

The example of discretion demonstrates the significant shift enabled by the introduction of new and emerging technologies into how organizations, such as the judiciary, function which is much more profound than a 'technological upgrade'. Yet, the foregoing also illustrates that a focus on accountability structures that try to adapt principles around technology rather than examining what is happening to fundamental principles, like discretion, may mask the size and nature of that shift.

3. Layering in Actor-Focused Accountability

Treating the incorporation of new and emerging technologies as a governance choice may be more revealing and thus facilitate greater recognition and scrutiny into the

[11] See Asaro, 'On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-making', 94 *International Review of the Red Cross* (2012).

[12] Citron, 'Technological Due Process', 85 *Washington University Law Review* (2008) 1249, at 1271.

rationale and justification for such choices and a fuller recognition of their effect on underlying principles and recognized modes of governance. In the example of discretion, it would widen the debate out from ways in which to ensure humans are 'in' or 'on' the loop to a recognition that, even where this could be achieved, the way in which discretion operates has changed. This would then require analysis of whether and when a change in the position of discretion is legitimate and acceptable, thus directing the focus to accountability of the actor who made the governance choice to change the role of discretion.

These types of governance shifts are evident across a range of bodies with a public mandate, whether at the international or national level. For example, Fleur Johns uses the shift to the use of biometric registration by the UN Office for the High Commissioner for Refugees to discuss the ways in which technology can expand 'international law's and institution's capacity'.[13] She discusses two guides on refugee determinations produced in the 1990s which 'contemplate the use of computers … [but] in limited and mostly instrumental terms' within a wider determination process as 'complex, creative and intensely human'.[14] She contrasts these to a shift towards biometric registration that radically alters the nature of the decision-making process. Using this and other examples, she warns that '[e]xercises of international legal authority that cannot be understood, represented, or justified in recognizable terms may be prone to rejection'.[15] This point emphasizes that transparency and scrutiny of shifts in governance choices are not only important for accountability purposes but for the wider trust and confidence in organizations that carry out public functions.

Starting from the position of governance choices also enables a wider lens on the impact of the introduction of new and emerging technologies on the full mandate of the actor rather than only on the direct effects of the technology concerned. The role of technology in policing provides another example of the potential reshaping of the way a particular task or mandate is carried out. Significant concerns have been raised about the risks of predictive policing to privacy and the threats to discrimination and profiling of individuals and communities, where it is used. However, there are also questions regarding the deployment of predictive policing tools in an operational context, and how this will affect police officers' ability to engage with the public, and to draw on their own experience. The concern is that officers will become overly dependent on technology and will engage in 'tablet policing'[16] at the expense of community policing, for example.[17] Such a shift could introduce its own security and human rights implications as well as impact on trust in key public bodies like the police which, while not caused by the new and emerging technology itself, could result from the policy decision to prioritize that way of working. Where questions of accountability only

[13] Johns, 'Data, Detection and the Redistribution of the Sensible in International Law' 111 *American Journal of International Law* (2017) 57 at 58.

[14] *Ibid.*, at 80.

[15] *Ibid.*, at 59.

[16] I.e., following instructions delivered by their tablet devices.

[17] This potentially affects the ability to deliver policing by consent. For a discussion on policing by consent, referencing the Peel Principles, see UK Home Office, 'Definition of policing by consent', 10 December 2012, available at https://www.gov.uk/government/publications/policing-by-consent/definition-of-policing-by-consent.

focus on how to minimize risks to human rights where predictive policing is used, they risk missing the potential for 'organizational and systemic trade-offs'[18] that may also entail risks. These risks – including of mandate trade-offs – may be even greater where the introduction of technology requires investment at a time where many national and international organizations with public functions face significant budget cuts. This may mean that parts of an organization's mandate are reduced – or even cut – while technological capability is increased, again requiring scrutiny.

The foregoing analysis illustrates that the employment of technology not only has the potential to adversely impact rights but also potentially changes the way in which a key public organization functions, raising significant governance questions. Yet, because it is enabled by technology, the change in the fulfilment of the mandate is at risk of a lack of scrutiny.

This is not to mean that all shifts carry significant risks or actually change the methodology of a mandate. For example, Future Advocacy and the Wellcome Trust have noted that in the health context the use of artificial intelligence techniques such as automation 'may free up HCP [healthcare practitioners] time that is currently occupied by routine administrative tasks, allowing them to spend more time interacting with patients'.[19] Thus, the employment of technology may also (re)enable a preferred means of fulfilling a mandate. However, it should be noted that this example clearly contains technology to an assistive rather than displacement task. Moreover, the report notes that this shift may not be stable in that 'as the technology improves and more tasks become automatable, it is increasingly possible that fewer 'human practitioners' will be required to run healthcare systems worldwide', meaning that there would be fewer nurses to spend time 'interacting with patients'.[20] This again illustrates the importance of accountability for governance choices, including an accountability model that can predict how technology may distort and disrupt governance choices in the future.

Starting with the mandate and the actor is also a counter to the concern that principles laid out in global administrative law, but also the rule of law more broadly and human rights,[21] are at risk of becoming 'obsolete'. This is because they can require transparency and openness in the governance choices organizations propose to make and critically embed the principle of 'bidirectional communication', which Benvenisti frames as a central tool of accountability, through participation in the co-option and design choices of such organizations. As noted in the Future Advocacy and Wellcome Trust report, '[t]here is [currently] a risk that these technologies are developed without the input of patients and those who use them – that is, the people who will be most impacted by these technologies'.[22] 'Points of friction' are often available once the

[18] Sandvik, Jacobsen and McDonald, 'Do No Harm: A Taxonomy of the Challenges of Humanitarian Experimentation' 99(1) *International Review of the Red Cross* (2017) 319, at 324.

[19] Future Advocacy and Wellcome Trust, *Ethical, Social and Political Challenges of Artificial Intelligence in Health* (2018) available at: https://wellcome.ac.uk/sites/default/files/ai-in-health-ethical-social-political-challenges.pdf at 26.

[20] *Ibid.*, at 27.

[21] See Harlow, *supra* note 3.

[22] Future Advocacy, *supra* note 19, at 36.

decision to employ the technology has been made (thus containing the question to how the technology operates and mitigation measures, as discussed above). Although not a term of art, 'points of friction' are means by which the approach and assumptions of the responsible entity are tested and challenged, and alternative points of view are raised and addressed. However, these points are not currently available at the point of choice and analysis of the impact of new and emerging technologies on those directly affected as well as on the fulfilment of an organization's mandate as a whole. Yet, this is the point at which the 'culture of accountability' discussed by Benvenisti in his article has some of the greatest potential for bite.

4. Conclusion

This article concludes that more work is needed to situate the demands for technological or algorithmic accountability within a wider accountability framework of governance choice. It thus locates questions of accountability with the actors that employ technology as part of a wider accountability matrix that also needs to address the wider global power dynamics of technology companies and the specificities of how technologies function. Starting with the choices the actors are making to embed technology rather than only from the perspective of the technology provides a fuller accountability lens and also means for transparency and scrutiny. This is because it starts from a recognized position of how actors carry out their mandates, and interrogates why a shift is needed and how that advances rather than degrades the principles on which that mandate is based; and where it does not, whether that can be overcome. From an administrative law perspective, such decisions to employ technology should therefore be transparent and subject to review as well as open to participation, consultation and feedback through models such as 'bidirectional communication'.