To be published in M Dubber and F Pasquale (eds.)
The Oxford Handbook of AI Ethics, Oxford University Press (2019) forthcoming

AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing

Karen Yeung, Andrew Howes, and Ganna Pogrebna

'Finding ways of developing and deploying new technologies with a purpose restricted to supporting individual freedom and dignity as well as the basic constitutional settlement of constitutional democracies, namely democracy, rule of law and fundamental rights is the challenge of our time.'

Paul Nemitz (2018)
Principal advisor
European Commission's Directorate-General
for Justice and Consumers

1. Introduction

The number and variety of topics in this volume illustrate the width, diversity of content and vagueness of the boundaries of 'AI Ethics' as a domain of inquiry. Within this discourse, increasing attention has been drawn to the capacity of socio-technical systems that utilise data-driven algorithms to classify, to make decisions and to control complex systems, including the use of machine learning and large datasets to generate predictions about future behaviour (hereafter 'AI' systems[1]) in ways that may interfere with human rights. The recent Cambridge Analytica scandal revealed how unlawfully harvested Facebook data from millions of voters in the UK, the US and elsewhere enabled malign actors to engage in political micro-targeting through the use of AI-driven social media content distribution systems, thereby interfering with their right to free and fair elections and thus threatening the integrity of democratic processes. The increasing use of algorithmic decision-making ('ADM') systems to inform custodial and other decisions within the criminal justice process may threaten several human rights, including the right to a fair trial, the presumption of innocence and the right to liberty and security. Systems of this kind are now used to inform, and often to automate, decisions about an individual's eligibility and entitlement to various benefits and opportunities, including housing, social security, finance, employment and other life-affecting opportunities, potentially interfering with rights of due process and rights to freedom from unfair or unlawful discrimination[2]. Because these systems have the capacity to operate both

[1] See European Commission (2018) "Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe." Available at https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe for more detail.

[2] Wagner, B. (2017) *Study On The Human Rights Dimensions of Automated Data Processing Techniques (In Particular Algorithms) And Possible Regulatory Implications'*. Council of Europe, Committee of Experts on internet intermediaries (MSI-NET). Available at https://rm.coe.int/study-hr-dimension-of-automated-data-processing-incl-algorithms/168075b94a (Accessed 6.11.18).

automatically and at scale, their capacity to affect thousands if not millions of people at a stroke can now occur at orders of magnitude and speeds not previously possible.[3]

This chapter has two overarching aims. Firstly, we argue that the international human rights framework provides the most promising set of standards for ensuring that AI systems are ethical in their design, development and deployment. Secondly, we sketch the basic contours of a comprehensive governance framework, which we call 'human rights-centred design, deliberation and oversight', for ensuring that AI can be relied upon to operate in ways that will not violate human rights. Four features of on-going discussions provide important context for our argument. First, the rubric of 'AI Ethics' is now used to encapsulate a multiplicity of value-based, societal concerns associated with the use of AI applications across an increasingly extensive and diverse range of social and economic activities. Secondly, there is a notable lack of clarity about the content of the normative values and principles that constitute the relevant 'ethical' standards to which AI should adhere . Thirdly, industry self-regulation is the predominant approach for bringing about 'ethical AI', reflected in a litany of 'ethical codes of conduct' promulgated by individual tech firms and various tech industry consortia published in response to the recent 'Tech Lash'.[4] These codes presuppose that the tech industry can formulate appropriate ethical norms for AI, and can be trusted to ensure that AI systems will duly adhere to those standards. Suggestions that more conventional regulation, involving legally mandated regulatory standards and enforcement mechanisms, are swiftly met by protest from industry that 'regulation stifles innovation.'[5] These protests assume that innovation is an unvarnished and unmitigated good, an unexamined belief that has resulted in technological innovation (particularly in the digital services industry) entrenching itself as the altar upon which cash-strapped contemporary governments worship, naïvely hoping that digital innovation will create jobs, stimulate economic growth and thereby fill diminishing governmental coffers left bare after propping up the banking sector which teetered on the brink of collapse following the global financial crisis in 2008. Fourthly, discussion of the need for meaningful *enforcement* of ethical standards is almost entirely absent from these initiatives.[6]

This paper proceeds in three stages. First, we argue that international human rights standards offer the most promising basis for developing a coherent and universally recognised set of standards that can be applied to meet many (albeit not all) of the normative concerns currently falling under the rubric of 'AI Ethics'. Second, the paper outlines the core

---

[3] Yeung, Karen (2019) *A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework*. Council of Europe MSI-AUT committee study (revised draft). Available at https://rm.coe.int/a-study-of-the-implications-of-advanced-digital-technologies-including/168094ad (Accessed 7.6.19).

[4] The Economist (2018). 'The techlash against Amazon, Facebook and Google - and what they can do'. Available at https://www.economist.com/briefing/2018/01/20/the-techlash-against-amazon-facebook-and-google-and-what-they-can-do (Accessed 6.11.18).

[5] There have, however, been more recent concessions by Big Tech about the need for legal regulation: e.g., Microsoft's call for legal regulation of facial recognition technology by states, and Facebook's Mark Zuckerberg's recent acknowledgement that some kind of regulation is needed in relation to data-driven social media content distribution systems.

[6] With the exception of recent legislation and proposed legislation (eg. in Germany, France and the UK) concerned with reducing the prevalence of extremist and terrorist media content online.

elements of a 'human-rights centred design, deliberation and oversight' approach to the governance of AI, explaining why such an approach is needed. Because much more theoretical and applied research is required to flesh out the details of our proposed approach, the third section sets out an agenda for further research, identifying the multiple lines of inquiry that must be pursued to develop the technical and organisational methods and systems that will be needed, based on the adaptation of existing engineering and regulatory techniques aimed at ensuring safe system design, re-configuring and extending these approaches to secure compliance with a much wider and more complex set of human rights norms. The fourth and final section concludes, reflecting on the limitations of our proposed approach and the magnitude of the challenges associated with making it implementable in real world settings. Nevertheless, we suggest that a 'human rights-centred design, deliberation and oversight' approach to the governance of AI offers a concrete proposal capable of delivering *genuinely* 'ethical AI', for at least four reasons, and to which we now turn.

2. Why should human rights lie at the core of 'AI ethics'?

Within contemporary discussions of 'AI ethics', there is no agreed set of ethical standards that should govern the operation of AI, reflected in the variety of ethical standards espoused in various voluntary 'AI ethics codes' that have emerged in recent years. The salience of 'AI ethics' reflects welcome recognition by the tech industry and policy-makers that AI systems may have significant adverse impacts,[7], with some values commonly appearing in these discussions, particularly those of 'transparency', 'fairness' and 'explainability'[8]. Yet the vagueness and elasticity of the scope and content of 'AI ethics' has meant that it currently operates as an empty vessel into which anyone (including the tech industry, and the so-called Digital Titans) can pour their preferred 'ethical' content. Without an agreed framework of norms that clearly identifies and articulates the relevant ethical standards which AI systems should be expected to comply with, little real progress will be made towards ensuring that these systems are in practice designed, developed and deployed in ways that will meet widely accepted ethical standards. Although there is scope for reasonable disagreement concerning what 'ethical conduct' requires in any given case, a core set of agreed norms that constitute the basic minimum below which conduct cannot fall if it can be appropriately characterised as ethically acceptable must be identified.[9]

---

[7] Yeung, Karen (2019) *A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework*. *Council of Europe MSI-AUT committee study (revised draft)*. Available at https://rm.coe.int/a-study-of-the-implications-of-advanced-digital-technologies-including/168094ad (Accessed 7.6.19).

[8] See Greene, Daniel, Anna Lauren Hoffmann, and Luke Stark (2019) "Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning." In *Proceedings of the 52nd Hawaii International Conference on System Sciences*; Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge et al. (2018) "AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations." *Minds and Machines* 28: 689-707.

[9] Nemitz, P. (2018). 'Constitutional Democracy and Technology in the Age of Artificial Intelligence'. *Phil Trans. A*. 376.

We believe that international human rights standards offer the most promising set of ethical standards for AI, as several civil society organisations have suggested.[10] As an international governance framework, human rights law is intended to establish global standards ('norms') and mechanisms of accountability that specify the way in which individuals are entitled to be treated. The UN Universal Declaration of Human Rights (UNHR) 1948, is perhaps the most well-known international human rights charter, based on a commitment that the appalling treatment of individuals that occurred during WWII should not only be condemned and prohibited outright, but ought never be repeated. Despite the number of, and variation between, regional human rights instruments in the Americas, Africa and Europe, and enshrined in the constitutions of individual nation states, they are all grounded on a shared commitment to uphold the inherent human dignity of each and every person, in which each individual is regarded of equal dignity and worth, wherever situated.[11] These shared foundations reflect the status of human rights standards as *basic moral entitlements* of every individual in virtue of their humanity, whether or not those entitlements are given explicit legal protection.[12]

The extent to which governments recognise these basic moral entitlements as *legally enforceable* rights varies considerably, partly due to differences in political ideology. In contemporary liberal democratic states, human rights are now widely recognised as essentially 'constitutional' in status to provide effective guarantees that individual freedom will be cherished and respected. In particular, the European Union's legal order is rooted in constitutional commitments to human rights, democracy and the rule of law, the so-called 'constitutional triumvirate' that form the foundational principles upon which political systems characterised as liberal constitutional democracies ultimately rest.[13] This brings us to a second reason why human rights norms provide the appropriate norms for securing 'ethical

---

10 See various reports by civil society organisations concerned with securing the protection of international human rights norms, e.g., Latonero, M (2019) *Governing Artificial Intelligence: Upholding Human Rights and Human Dignity*, Data & Society. Available at https://datasociety.net/wpcontent/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf (Accessed 6 May 2019). *The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning Systems* (2018) Available at https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/; *The Montreal Declaration for a Responsible Development of Artificial Intelligence: A Participatory Process* (2017). Available at https://nouvelles.umontreal.ca/en/article/2017/11/03/montreal-declaration-for-a-responsible-development-of-artificial-intelligence/; Access Now (see https://www.accessnow.org/tag/artificial-intelligence/ for various reports); Data & Society (see https://datasociety.net/); IEEE (see https://www.ieee.org/) report on ethically aligned design for AI (Available at https://ethicsinaction.ieee.org/) which lists as its first principle that AI design should not infringe international human rights; AI Now Report (2018). Available at https://ainowinstitute.org/AI_Now_2018_Report.pdf. See also McGregor, L et al (2019) 'International Human Rights Law as a Framework for Algorithmic Accountability' 68 *ICLQ* 309-343.

11 Latonero, M (2019) *Governing Artificial Intelligence: Upholding Human Rights and Human Dignity*, Data & Society at p. Available at https://datasociety.net/wpcontent/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf (Accessed 6 May 2019).

12 Dworkin, R. (1977). *Taking Rights Seriously*. London, Duckworth.

13 Nemitz, P. (2018). 'Constitutional Democracy and Technology in the Age of Artificial Intelligence'. *Phil Trans. A*. 376.

AI': because a commitment to effective human rights protection is part and parcel of democratic constitutional orders. In a world in which AI systems increasingly configure our collective and individual environments, entitlements and access to, or exclusion from, opportunities and resources, it is essential that the protection of human rights, alongside respect for the rule of law and the protection of democracy, is assured to maintain the character of our political communities as constitutional democratic orders in which every individual is free to pursue his or her own version of the good life as far as this is possible within a framework of peaceful and stable cooperation framework underpinned by the rule of law. [14] This contrasts starkly with most contemporary AI ethics codes, which typically outline a series of 'ethical' principles that have been effectively plucked out of the air, without any grounding in a specific vision of the character and kind of political community that it is committed to establishing and maintaining, and which those principles are intended to secure and protect.[15]

The well-developed institutional framework through which systematic attempts are made to monitor, promote and protect adherence to human rights norms around the world provide two additional reasons in support of adopting human rights standards to ensure the ethical governance of AI. Despite considerable variation in the range and scope of rights enumerated in formal Charters of Rights, there is a well-established analytical framework through which tension and conflict between rights, and between rights and collective interests of considerable importance in democratic societies, are resolved in specific cases through the application of a structured form of reasoned evaluation. This approach is exemplified in the structure and articulation of human rights norms within the European Convention of Human Rights (the 'ECHR'). The ECHR (ratified by 47 countries) specifies a series of human rights norms, including (among others) the right to freedom of expression, the right to life, the right to private and home life, the right to freedom of assembly and religion, for example, all of which must be guaranteed to all individuals and effectively protected. However, for many of those rights, certain qualifications are permitted, in order to ensure respect for a narrow range of clearly specified purposes that are necessary in a democratic society, and provided that any such qualifications are prescribed by law and proportionate in relation to those purposes. So, for example, Article 10 of the ECHR provides that:

> (1) Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers. This Article shall not prevent States from requiring the licensing of broadcasting, television or cinema enterprises.
>
> (2) The exercise of these freedoms, since it carries with it duties and responsibilities, may be subject to such formalities, conditions, restrictions or penalties as are prescribed by law and are necessary in a democratic society, in the interests of national security, territorial integrity or public safety, for the prevention of disorder or crime, for the protection of health or morals, for the protection of the reputation or rights of others, for preventing the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary.

---

[14] Hildebrandt, M. (2015) *Smart technologies and the end (s) of law: novel entanglements of law and technology*. Edward Elgar Publishing.

[15] See for example the *Beijing AI Principles*. Available at https://www.baai.ac.cn/blog/beijing-ai-principles (Accessed 12.6.19)

Accordingly, although freedom of expression is essential to ensure, amongst other things, free democratic debate and individual self-determination, legal restrictions on expression may be permissible for the purposes specified in Article 10(2). Accordingly, restrictions on expression may be justified in order, for example, to protect individual rights to privacy or the right to free and fair elections, if they conflict in particular cases, provided that restrictions are legally prescribed and go no further than the minimum necessary to protect these rights.

This structured framework for reasoned resolution of conflict arsing between competing rights and collective interests in specific cases is widely understood by human rights lawyers and practitioners, forming an essential part of a 'human rights approach'. This framework overcomes another shortcoming in existing codes of ethical conduct: their failure to acknowledge potential conflict between ethical norms, and the lack of any guidance concerning how those conflicts will or ought to be resolved in the design and operation of AI systems. Of the codes which do acknowledge potential conflict , little is offered by way of guidance concerning how to resolving such conflict: both in the codes themselves, and in much of the on-going 'AI ethics' literature, beyond suggesting one should seek help from an ethics expert[16].

In contrast, the well-established human rights approach to the resolution of ethical conflict is informed by, and developed through, a substantial body of authoritative rulings handed down by judicial institutions (at both international and national level) responsible for adjudicating human rights complaints. These adjudicatory bodies, which determine allegations of human rights violations lodged by individual complainants, form part of a larger institutional framework that has developed over time to monitor, promote and protect human rights, and includes a diverse network of actors in the UN system, other regional human rights organisations (such as the Council of Europe and a wide range of civil society organisations focused on the protection of human rights), national courts and administrative agencies, academics and other human rights advocates. The institutional framework for rights monitoring, oversight and adjudication provides a further reason why human rights norms provide the most promising basis for AI ethics standards. The dynamic and evolving corpus of judicial decisions can help elucidate the scope of justified interferences with particular rights in concrete cases, offering concrete guidance to those involved in the design, development and implementation of AI systems concerning what human rights compliance requires. Most importantly, perhaps, these human rights norms are both internationally recognised and, in many jurisdictions, supported by law, thereby providing a set of national and international institutions through which allegations of human rights violations can be investigated and enforced, and hence offer a means for real and effective protection.

This contrasts sharply with the prevailing self-regulatory model favoured by the tech industry and to which most national and regional governments (including the EU) have acquiesced[17].

---

[16] See Council of Europe, "Guidelines on Artificial Intelligence and Data Protection", January 2019, https://rm.coe.int/guidelines-on-artificial-intelligence-and-data-protection/168091f9d8. While the EU High Level Expert Group's *Ethics Guidelines for Trustworthy AI* (2019) refers to the need for reasoned evaluation, transparency and documentation of how such ethical trade-offs are resolved when they are encountered by those involved in designing, developing and deploying AI systems, it offers no substantive guidance concerning how that evaluation should be conducted.

[17] Self-regulation is the controlling of a process or activity by the people or organisations that are involved in it rather than by an outside organisation such as government. Voluntary self-regulation or 'pure' self regulation entails the private firm or industry making and enforcing the rules,

Although self-regulation has been effective in a handful of industries which can be understood as a 'community of shared fate'[18] (e.g. USA nuclear industry after Three Mile Island via INPO), there are good reasons to doubt their effectiveness in general,[19] given that self-regulatory standards that have no legally binding force.[20] Because tech firms operate in highly competitive global markets in which securing the first mover advantage is often accompanied by the capacity to reap the extensive benefits arising from global network effects (eg Google Maps) it is naïve to expect that they can be trusted to abide by voluntary standards when faced with such powerful commercial imperatives. It is hardly surprising that critics have dismissed these voluntary codes of conduct as 'ethics washing'[21] given overwhelming evidence that the tech industry cannot be relied upon to honour its voluntary commitments.[22] Nemitz describes the growth of these initiatives as a 'genius move' by the tech industry, allowing the industry to focus attention and resources on the ethics of AI to delay the debate and work on the *law* for AI.[23] As Hagendorff comments:

> "AI ethics – or ethics in general – lacks mechanisms to reinforce its own normative claims. Of course, the enforcement of ethical principles may involve reputational losses….Yet…these mechanisms are rather weak and pose no eminent (sic) threat….Ethical guidelines of the AI industry serve to suggest to legislators that internal self-governance is sufficient, and that no

independent of direct government involvement: N Gunningham (2011) 'Investigation of Industry Self-Regulation in Workplace Health and Safety in New Zealand. Available at http://regnet.anu.edu.au/sites/default/files/publications/attachments/2015-04/NG_investigation-industry-self-regulation-whss-nz_0.pdf (Accessed 14.6.2019).

18 J Rees (1994). *Hostages to Each Other: The Transformation of Nuclear Safety Since Three Mile Island,* Chicago: University of Chicago Press.

19 See OECD (2015) *Industry Self-Regulation: Role and Use in Supporting Consumer Interests.* DSTI/CP(2014)4/FINAL 20-21 for a list of shortcomings. Regulation expert Neil Gunningham observes that, "The extent to which self-regulation in practice has either positive or negative attributes will depend very much on the social and economic context within which it operates and on the particular characteristics of the scheme itself. Nevertheless it is fair to say that 'pure' self-regulation is rarely effective in achieving social objectives because of the gap between private industry interests and the public interest." See Gunningham, *supra* n 21,3.

20 Borgesius, Frederik Zuiderveen (2018) *Discrimination, Artificial Intelligence and Algorithmic Decision-Making.* Council of Europe, Directorate General for Democracy. Available at https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73, (accessed June 3, 2019).

21 Wagner, B. (2018) 'Ethics as an Escape from Regulation: From ethics-washing to ethics-shopping?' in E Bayamlioglu et al (eds.) *Being Profiled: Cogitas Ergo Sum.* Amsterdam University Press, Amsterdam.

22 For a sobering account of Facebook's repeated failure to honour its publicly stated commitments, see UK House of Commons, Digital Culture Media and Sports Committee, *Disinformation and 'fake news' : Final Report*, Eighth Report of Session 2017-2019, 14 February, HC 1791. Available at https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/1791/1791.pdf (Accessed 6.5.19).

23 Nemitz, P. (2018). 'Constitutional Democracy and Technology in the Age of Artificial Intelligence'. *Phil Trans. A*. 376.

> specific laws are necessary to mitigate possible technological risks and to eliminate scenarios of abuse …. And even when more concrete laws concerning AI systems are demanded, as recently done by Google…these demands remain relatively vague and superficial."[24]

Of the handful of ethical AI proposals advocated by civil society and other international organisations that have drawn attention to the need to ensure that AI systems respect human rights norms, they have paid scant attention to their enforcement.[25]

3. Why 'human rights-centred design, deliberation and oversight' of AI?

The ineffectiveness of the prevailing self-regulatory approach to 'ethical AI' demonstrates that an alternative governance model is needed: one that is (1) anchored in human rights norms and a human rights approach (2) utilises a coherent and integrated suite of technical, organisational and evaluation tools and techniques, that is (3) subject to legally mandated external oversight by an independent regulator with appropriate investigatory and enforcement powers, and (4) provides opportunities for meaningful stakeholder and public consultation and deliberation. In the next section, we develop a governance framework for AI intended to do just that, which we call 'human rights-centred design, deliberation and oversight'. Although much more foundational work remains to be done, both to specify the content and contours of this approach more fully and to render it capable of practical implementation, we believe the our proposed framework offers a concrete approach that can bring an end to 'ethics washing' by securing the design, development and deployment of human rights-compliant AI systems in real world settings. The core elements of our approach are outlined in the following discussion.

3.1 What is 'human rights-centred design, deliberation and oversight'?

Our proposed governance regime for AI and other relevant automated systems (understood as complex socio-technical systems which includes the data upon which they rely for training and operation) seeks to ensure that these systems will be human-rights compliant and reflect

---

[24] Hagendorf, T. (2019) 'The Ethics of AI Ethics: An Evaluation of Guidelines' at 1. Available at https://arxiv.org/abs/1903.03425 (Accessed 6 May 2019). As the recent EU 'Algo-Aware' (Dec 2018) project observes, "Across the globe, the majority of initiatives (ie concerned with programmes aimed at securing algorithmic accountability) are very recent or still in development. Additionally, there are limited concrete legislative or regulatory initiatives being implemented." European Commission's Directorate-General for Communications Networks, Content and Technology (2019) Algo-Aware, *State of the Art Report: Algorithmic Decision-Making*. Available at https://www.algoaware.eu/state-of-the-art-report/ (Accessed 7.6.19)

[25] For example, The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning Systems, which focuses only on rights to equality and non-discrimination in ML systems, draws attention to the character of human rights as a "universally ascribed system of values based on the rule of law" that constitute a "universally binding, actionable set of standards" (per paragraph 9) for which "prompt and effective remedies" must be available against "those responsible for violations" (per paragraph 49): See https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/for more detail. Although the EU's *Ethical Guidelines for Trustworthy AI* places human rights protection at their foundation, offering an 'Assessment List' in order to provide concrete guidance to tech firms seeking to adhere to the ethical guidance thereby provided, the guidelines remain entirely voluntary, and make no provision for external, independent oversight and enforcement.

the core values that underpin the rule of law.[26] It entails systematic consideration of human rights concerns at every stage of system design, development and implementation (making interventions where this is identified as necessary). Such a regime should be mandated by law, subject to external oversight by independent, properly resourced regulatory authorities with appropriate powers of investigation and enforcement, and which provides for input from both technical and human rights experts, on the one hand, and meaningful input and deliberation from affected stakeholders and the general public on the other. Our approach seeks to integrate *both* ethical design strategies, technical tools and techniques for software and system design, verification, testing and auditing, together with social and organisational approaches to effective and legitimate governance. In so doing, our approach seeks to integrate a range of methods from a wide variety of intersecting disciplinary perspectives, including:

(a) pragmatic 'ethics in design' frameworks developed by applied ethicists concerned with ensuring that due attention is given to moral values in technical innovation processes early on in the technical design process, with the aim of integrating values and engineering design.[27] For example, the 'Value-Sensitive Design (VSN)' approach developed by Friedman and others builds on the insights of the human-computer interaction community, with a concern to incorporate human and moral values into the design of information technology, connecting those who design systems with those affected by them and other stakeholders.[28]

(b) engineering techniques concerned with 'hard-wiring' specific values into a socio-technical system's design and operation. Although the most established methods and experience have hitherto focused on ensuring the value of safety, primarily via safety engineering techniques, an increasing body of work in software design and engineering has expanded the range of values which engineers have sought to encode and protect via system design. These include privacy (referred to as 'privacy enhancing technologies' or 'privacy by design') security ('security by design') and, more recently, data protection principles ('data protection by design'). In the realm of machine learning, a growing body of technical research in 'explainable AI' (XAI) has been devoted to enhancing the capacity for machine learning

---

[26] Our proposed approach to AI governance can be understood as compatible with, and complementary to, Hildebrandt's concept of 'legal protection by design (LPbD).' Hildebrandt's LPbD places greater emphasis on articulating the challenges for legal protection and the rule of law posed by code-driven technologies (including an identification of the substantive and procedural opportunities and capacities which computational systems must provide in order to ensure that the values and commitments underpinning the rule of law are reflected in these systems, such as rights of contestation, rights to demand an explanation and justification etc). In contrast, our 'human rights-centred design, deliberation and oversight places greater emphasis on developing concrete legal, technical and organisational governance methods and techniques for ensuring that human rights protection is implemented into complex socio-technical systems that utilise AI technologies: See M Hildebrandt (2015) *Smart Technologies and the End(s) of Law.* Edward Elgar, Chelteham. M Hildebrandt (2019) *Law for Computer Scientists*, Oxford University Press. Available at https://lawforcomputerscientists.pubpub.org (Accessed 19.6.19)

[27] van den Hoven, J., Miller, S and Pogge, T (eds.) (2017). *Designing in Ethics*. Cambridge, Cambridge University Press: 28.

[28] Friedman, B (1996) 'Value-Sensitive design.' *Interactions* 3:16-23; Friedman, B., Kahn, P., and Borning, A. (2002) *Value-Sensitive Design: Theory and Methods*. CSE Technical Report 02-12-01. Seattle: University of Washington.

systems to provide avenues through which humans can better understand the logic by which machine learning systems generate outputs as well as techniques for improving the 'fairness, accountability and transparency' (FAT) of these outputs by seeking to identify and eliminate unfair discrimination. Taken together, these techniques can be understood as falling within this expanding family of technical approaches to securing ethical values beyond that of safety;

(c) a suite of methods and techniques in software design and engineering, including various forms of assessment, testing and evaluation to identify whether particular aspects of a system (provably) meet certain pre-specified standards and requirements;[29]

(d) organisational accountability mechanisms and established regulatory techniques used in safety critical domains that that operate ex ante, requiring systematic evaluation of safety concerns and appropriate interventions *before* a system is deployed, as well as ex post techniques that apply after the system has been deployed and which have been designed and developed to ensure the traceability and auditability of system behaviour via systematic recording and logging of a system's design and operation and any alterations thereto;[30]

(e) a range of regulatory governance techniques that have been used effectively in other contexts, including the use of (i) impact assessment' tools and methods that incorporate opportunities and mechanisms for facilitating stakeholder consultation, engagement and deliberation, particularly in relation to the design, development and deployment of 'high risk' applications[31]; (ii) risk-based approaches to regulation that seek to ensure that 'high risk' systems are subject to the most intensive and demanding scrutiny, whilst the burdens of demonstrating human rights compliance for low risk systems are proportionately less demanding[32] (iii) 'meta-regulatory' approaches that seeks to harness the knowledge and expertise within firms themselves in the service of regulatory compliance, overseen by a public regulator endowed with powers of investigation and sanction, and (iv) post-implementation monitoring ('AI system vigilance'), in order to systematically and transparently track adverse events in order to identify problems and failures as early as possible in order facilitate swift corrective interventions.

Our expectation is that these frameworks and methods can be adapted and refined to ensure that respect for *human rights norms* is integrated into system design, while incorporating a human rights approach to the resolution of conflict and tension between human rights norms, or between human rights norms and important collective interests, which may arise in specific contexts and circumstances. At the same time, we anticipate the need for new techniques and frameworks to accommodate novel human rights risks that the development

---

29 Kroll, J. A., et al (2016) "Accountable algorithms." *U. Pa. L. Rev.* 165: 633.

30 Rieke, A., Bogen, M. and Robsinson, D.G. (2018) *Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods*, An Upturn and Omidyar Network Report. Available at https://www.omidyar.com/insights/public-scrutiny-automated-decisions-early-lessons-and-emerging-methods (Accessed 3.6.19)

31 See, for example, data protection impact assessment eg Mantelero, A. (2018) 'AI and Big Data: A blueprint for a human rights, social and ethical impact assessment.' *Computer Law & Security Review* 34(4): 754-772 and human rights impact assessment, eg Raso, F.A. et al. (2018) *Artificial Intelligence & Human Rights: Opportunities & Risks*. Berkman Klein Center Research Publication.

32 Black, J. (2008). *Risk-Based Regulation: Choices, Practices and Lessons Being Learned. Risk and Regulatory Policy*. OECD: Paris.

and deployment of AI systems may generate. For example, these are likely to include new governance frameworks and oversight mechanisms to ensure that data-driven experimentation on human users when undertaken outside conventional academic research settings is undertaken in a human-rights compliant manner, while the necessity of 'in the wild testing' of AI systems generate novel governance challenges that do not arise in circumstances where product and service development phase can be sharply delineated from their deployment.

3.2 Core Principles of Human Rights-Centred Design, Deliberation and Oversight

The methods and techniques listed above vary widely in their disciplinary foundations and in the original contexts of their development. Our proposal seeks to draw them together in an integrated manner, appropriately adapted towards ensuring conformity with human rights norms, as the basis for a comprehensive design and governance regime constructed around the following four core principles in which human rights norms provide the foundational ethical standards which AI systems must demonstrably comply with:

1. Design and deliberation
2. Assessment, testing and evaluation
3. Independent oversight, investigation and sanction
4. Traceability, evidence and proof

Each of these principles is briefly outlined below.

Principle 1: Design and deliberation

Central to our approach is a requirement that AI systems should be designed and configured to operate in ways that are compliant with universal human rights standards (such as those, for example, set out in the ECHR), and that, at least for systems identified during the design and development phase as posing a 'high risk' of interfering with human rights, affected stakeholders are consulted about the proposal and given opportunities to express their views about the proposed system's potential impact, in discussion with the system's designers. Consultation with affected stakeholders and the general public during the initial phases of system design contributes to the overall legitimacy of the regime, understood in terms of respect for democratic values and affected communities, and should help system designers to identify which aspects of the system's design and proposed operation need reconsideration. Where the risks to human rights are assessed as 'high' or 'very high'[33] this would trigger an obligation on system designers to reconsider and redesign the system and/or proposed business model[34] in order to reduce those risks to a form and level regarded as tolerable (understood in terms of a human rights approach to the resolution of conflict between rights and collective interests), in ways that duly accommodate concerns expressed by affected

---

[33] Jasanoff, S. (2016) *The ethics of invention: technology and the human future*. WW Norton & Company. Raso, F. A. et al, (2018) *Artificial Intelligence & Human Rights: Opportunities & Risks*. Berkman Klein Center Research Publication.

[34] On the potential discriminatory impact of data-driven business models, see Ali, M., et al (2019) 'Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes.' Retrieved from https://arxiv.org/pdf/1904.02095.pdf.

stakeholders and in recognition of the individual and collective benefits which the system is expected to generate.[35]

Principle 2: Assessment, testing and evaluation

Users and others affected by the operation of AI systems (including the general public) can only have justified confidence that AI systems do in fact comply with human rights standards if these systems can be subjected to formal assessment and testing to evaluate their compliance with human rights standards, and if these occur regularly throughout the entire lifecycle of system development: from the initial formulation of a proposal through to design, specification, development, prototyping, and real world implementation, and which includes periodic evaluation of the data sets upon which the system has been trained and upon which it operates.[36]

These evaluations form a core element of an overarching 'human rights risk management' approach, which aims to identify potential human rights risks *before* the deployment of AI and other relevant automated systems, and which occurs within a larger 'meta-regulatory' approach to AI governance in which AI system developers and owners are subject to legal duties to demonstrate to a public regulatory authority that their system is human rights compliant.[37] If significant risks to human rights compliance are identified, system developers must reconsider the design specification and system requirements with a view to modifying them in order to reduce those risks to a level that satisfies the tests of necessity and proportion[38] – or, in cases where the threats to human rights are disproportionate and thus unacceptably high, to refrain from proceeding with the development of the system in the

---

35 The participatory approach to social impact assessment referred to in the Council of Europe's AI Guideline strongly resonates with the role that our approach ascribes to public deliberation: see The Council of Europe, Guidelines on AI, (19 Feb 2019) https://rm.coe.int/guidelines-on-artificial-intelligence-and-data-protection/168091f9d8 at 23-24.

36 Kroll, Joshua A., Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. "Accountable algorithms." *U. Pa. L. Rev.* 165 (2016): 633; Borgesius, Frederik Zuiderveen *Discrimination, Artificial Intelligence and Algorithmic Decision-Making,* Council of Europe, Directorate General for Democracy at 51. Available at https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73, accessed 3.6.2019; Rieke, A., Bogen, M. and Robsinson, D.G. (2018) *Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods*, An Upturn and Omidyar Network Report. Available at https://www.omidyar.com/insights/public-scrutiny-automated-decisions-early-lessons-and-emerging-methods (Accessed 3.6.19).

37 Also called 'management-based' regulation and 'enforced self-regulation', meta-regulation refers to a strategy in which regulators do not prescribe how regulated firms should comply, but instead require them to develop their own systems for compliance with legally mandated goals and to demonstrate that compliance to the regulator: Black, J. (2012) 'Paradoxes and Failures: "New Governance" Techniques and the Financial Crisis." *Modern Law Review* 75: 1037, 1045-1048.

38 The formulation of the appropriate legal standard would need to reflect the established proportionality assessment that is well-established in addressing human rights conflicts and conflicts between human rights and legitimate collective interests, operating as the human rights equivalent as the 'as low as reasonably practicable' ('ALARP') requirement that applies to legal duties to ensure the safety of complex systems, per Hopkins, A. (2012) "Explaining Safety Case". Regulatory Institutions Network Working Paper 87. Available via SSRN network; Thomas, M. (2017). *Safety Critical Systems. Gresham Lectures.* London.

form proposed. Once the system has been implemented, periodic review must be undertaken and test and assessment documents duly filed with the public authority. A system of 'AI vigilance' is also needed, entailing the systematic recording of adverse incidents arising from system operations, including potential human rights violations reported by users or the wider public, triggering an obligation on the system provider to review and reassess the system's design and operation, and to report and publicly register any modifications to the system undertaken following this evaluation. Systematic and periodic post-implementation monitoring and vigilance is needed to ensure that AI systems continue to operate in a human-rights compliant manner because, once implemented into real-world settings, AI systems will invariably display emergent effects that are both difficult to anticipate, and may scale very rapidly. Accordingly, there is also an accompanying need for more systematic and sustained research concerned with modelling social systems, in order to better anticipate and predict their unintended adverse societal effects.

Principle 3: Independent oversight, investigation and sanction

In order to provide meaningful assurance that AI systems are in fact human rights compliant, rather than merely *claiming* to be human rights-compliant, independent oversight by an external, properly resourced, technically competent oversight body invested with legal powers of investigation and sanction is essential.[39] Because the operation of market-forces cannot provide those who design, develop and deploy AI systems with sufficient incentives to invest the required resources necessary to ensure that AI systems are human rights compliant, our proposed approach must operate within a *legally mandated* institutional structure, including an oversight body with a duty to monitor and enforce substantive and procedural (regulatory) requirements, including those concerning robust design, verification, testing, and evaluation (including appropriate documentation demonstrating that these requirements have been fulfilled) supported by legally mandated stakeholder and public consultation where proposed AI systems pose a 'high risk' to human rights.

We suggest that independent oversight is best designed within a meta-regulatory framework, in which legal duties are placed on AI system developers and operators to demonstrate to a public authority that their systems are human rights compliant.[40] Although there a wide variety of approaches that can be understood as meta-regulatory in form[41] the so-called 'safety case', properly implemented, is considered to have significantly contributed to ensuring the safety of complex systems in several domains, including safety regulation for off-shore petroleum drilling through to the regulation of workplace safety adopted in several

---

39 Borgesius, Frederik Zuiderveen (2018) *Discrimination, Artificial Intelligence and Algorithmic Decision-Making,* Council of Europe, Directorate General for Democracy, Available at https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73, (Accessed 3. 6.19).

40 See n.37 above.

41 See Black, J. (2006). 'Managing Regulatory Risks and Defining the Parameters of Blame: A Focus on the Australian Prudential Regulation Authority.' *Law and Policy* 28: 1-26; Gilad, S. (2010). 'It runs in the family: Meta-regulation and its siblings.' *Regulation & Governance* 4: 485-506. Coglianese, C. and E. Mendelson (2010). Meta-Regulation and Self-Regulation. In R. Baldwin, C. Hood and M. Lodge. (eds.) *The Oxford Handbook of Regulation*. New York, Oxford University Press**:** 146-168.

Anglo-Commonwealth legal systems.[42] In his discussion of offshore petroleum drilling, Hopkins highlights five basic features of a safety case approach: (1) All operators must prepare a systematic risk (or hazard) management framework which identifies all major hazards and provides detailed plans for how these hazards will be managed, specifying the controls that will be put in place to deal with the identified hazards, and the measures that will be taken to ensure that controls continue to function as intended; (2) A requirement for the operator to 'make the case' to the regulator, that is, to demonstrate to the regulator that the processes that have been undertaken to identify hazards, the methodology they have used to assess risks, and the reasoning (and evidence) that has led them to choose one control rather than another, should be regarded as acceptable. It is then for the regulator to accept (or reject) the case. Although a safety case gives operators considerable independence and flexibility in determining how they will respond to hazards, they do not have free reign: thus if an operator proposes to adopt an inadequate standard, a safety case regulator may challenge the operator and require the adoption of a better standard; (3) A competent, independent and properly resourced regulator with the requisite level of expertise and who can engage in meaningful scrutiny. The regulator's role is not to ensure that hardware is working, or that documents are up to date, but to *audit against the safety case*, to ensure that the specified controls are functioning as intended, and this necessitates a sophisticated understanding of accident causation and prevention; (4) Employee participation, both in the development of safety cases, and with whom the regulatory officials carrying out site audits must consult; and (5) A general legal duty of care imposed on the operator to do whatever is reasonably practicable to identify and control all hazards. An operator cannot claim to be in compliance just because it has completed a hazard identification process. It is the general duty of care that raises a safety case regime above a 'tick box' or 'blind compliance' mentality, so that a hazard identification process that is demonstrably inadequate would fail to meet the requisite standard.[43]

---

[42] The so-called 'safety case' movement emerged in the early 1990s in both the UK and USA as an approach to safety certification involving approval and oversight of complex systems, such as aircraft, nuclear power plants and offshore oil exploration. A Hopkins (2012). 'Explaining the "safety case"', *Regulatory Institutions Network*, Working Paper 87. Available at http://www.csb.gov/assets/1/7/WorkingPaper_87.pdf. There have, however, been criticisms of a safety case approach, including concerns about problems of confirmation bias, the need to consider worst case scenarios, reliance on probabilistic assessment to provide assurances of safety rather than the opposite goal of identifying unrecognised hazards, and examples of highly successful process based (rather than performance-based) approaches to securing safety in relation to submarines (eg the SUBSAFE programme): see N Leveson (2011) 'The Use of Safety Cases in Certification and Regulation.' *MIT Engineering Systems Division Working Paper Series*. Available at http://sunnyday.mit.edu/SafetyCases.pdf (Accessed 12 June 2019) 7-9. Leveson observes that the British Health and Safety Executive has applied a safety case regime widely to UK industries, pursuant to which responsibility for controlling risks is placed primarily on those who create and manage hazardous systems, based on three principles: (a) those who create the risks are responsible for controlling those risks, (b) safe operations are achieved by setting and achieving goals rather than by following prescriptive rules, (c) while those goals are set out in legislation, it is for the system providers and operators to develop what they consider to be appropriate methods to achieve those goals.

[43] Although the general duty of care is linguistically quite imprecise, its meaning has been elaborated on via case law, through numerous cases in which courts have had to decide whether the duty has been complied with. This case law gives fairly clear guidance as to what the general duty means in particular cases: Hopkins, A (2012) "Explaining Safety Case". Regulatory Institutions Network Working Paper 87. Available via SSRN network;*Safety Science* 49: 110-120; Thomas, M. (2017). *Safety Critical Systems*. Gresham Lectures. London.

Regulatory regimes of this kind allow (although they do not necessitate) the possibility of ex ante licensing by a designated public authority in the case of particularly human rights sensitive, 'high risk' systems, such as facial recognition systems intended for use by governments to identify individuals of interest in public places.[44] Applying the underlying logic and structure of the 'safety case' approach to human rights compliance would provide developers with considerable flexibility in seeking to 'make the case' to the regulator to demonstrate that their proposed AI systems can be expected to operate in human rights compliant ways.

Principle 4: Traceability, evidence and proof

In order to facilitate meaningful independent oversight and evaluation, AI systems must be designed and built to secure auditability: this means more than merely securing transparency, but is aimed at ensuring that they can be subject to *meaningful review*, thus providing a concrete evidential trial for securing human accountability over AI systems.[45] Not only is it necessary that systems be constructed to produce evidence that they operate as desired,[46] there must be a legal obligation to do so, requiring that crucial design decisions, the testing/assessment process and the outcome of those processes, and the operation of the system itself, are properly documented and provide a clear evidence trail that can be audited by external experts. Drawing again on the experience of the 'safety case' approach , which entails imposing a legal duty on operators to demonstrate to the regulator that robust and comprehensive systems are in place that reduce safety risks to a level that is 'as low as reasonably practical', we envisage the imposition of a suitably formulated legal duty on AI systems developers, owners and operators to demonstrate that these systems are human rights compliant.

To discharge this legal duty, AI system developers would also be subject to legal duties to prepare, maintain and securely store system design documentation, testing and evaluation reports and the system must be designed to routinely generate operational logs which can be inspected and audited by an independent, suitably competent authority. Taken together, these provide an audit trail through which system designers and developers can demonstrate that they have undertaken human rights 'due diligence' - thereby discharging their legal duty to demonstrate that they have discharged their legal duty to reduce the risk of human rights violations to an acceptable level. These traceability and evidential requirements apply to both the design and development phase (including verification and validation requirements), and the operation and implementation of systems (logging and black box recording of system operations). Taken together, these obligations are intended to ensure that robust and systematic transparency mechanisms are put in place, the aim of which is not complete

---

44 See for example Big Brother Watch (2018) *Face Off – The Lawless Growth of Facial Recognition in UK Policing.* Available at https://bigbrotherwatch.org.uk/wp-content/uploads/2018/05/Face-Off-final-digital-1.pdf (Accessed 12.6.19).

45 Bryson, JJ and Theodorou, A (2019) 'How Society Can Maintain Human-Centric Artificial Intelligence.' In M. Toivonen-Noro, and E. Saari (eds.) *Human-Centered Digitalization and Services,* Springer 12-13.

46 Kroll J et al (2016) 'Accountable algorithms.' *U. Pa. L. Rev.* 165: 633.

comprehension, but to provide sufficient information to ensure that human accountability for AI systems can be maintained.[47]

This integrated approach to AI governance grounded on the above principles can be understood as a response to the Council of Europe's call for a 'human-rights oriented development of technology'.[48] As Alessandro Mantalero has claimed:

> "Innovation must be developed responsibly, taking the safeguard of fundamental rights as the pre-eminent goal…This necessarily requires the development of assessment procedures, the adoption of participatory models and supervisory authorities. A human rights-oriented development of tech might increase costs and force developers and business to slow their current time-to-market, as the impact of products and services on individual rights and society have to be assessed in advance. At the same time, in the medium to long-term, this approach will reduce costs and increase efficiency (eg more accurate prediction and decision systems, increased trust, fewer complaints). Moreover, businesses and societies are mature enough to view responsibility towards individuals and society as the primary goal in AI development"[49]

4. Getting from here to there: a research agenda

The four principles outlined in the previous section demand revision to many aspects of software engineering (SE) practice. While a suite of relevant engineering and regulatory governance techniques are already in use in some specific areas, they require significant adaptation and generalisation to support meaningful human rights evaluation and compliance. Changes to SE practice must be complemented by a focused human rights-centred design research agenda in computer science. Such an agenda would draw together the currently fragmented activity in relevant software engineering disciplines (including SE, Cyber security, HCI, Verification) and also consider their continued relevance to the software lifecycle of AI Systems in particular, which are likely to require new design processes. Rather than offer a detailed research agenda here, we offer instead a 'manifesto', which identifies and briefly outlines some of the technical, engineering and governance challenges that must be met if SE is to provide assurances of human rights compliance. Some of these topics are existing areas of practice in software engineering and others are established research disciplines. None, however, have any tradition of human rights-centred design and most are only just beginning to consider appropriate software engineering practice for AI systems.

4.1 Requirements analysis

In SE, requirements analysis concerns the identification of the needs to be met by a new software system. The commercial orientation and diversity of approaches to requirements analysis presents a severe challenge to Principle 1 of our human rights-centred design agenda, and is made more acute by the fact that requirements analysis has been developed for non-AI systems. Corporate practice is strongly oriented to identifying the requirements of

---

47 Bryson and Theodorou n 55 at 14.

48 Mantalero, A. (2018) *AI and Data Protection, Challenges and possible remedies.* Study for Council of Europe. Available at https://rm.coe.int/artificial-intelligence-and-data-protection-challenges-and-possible-re/168091f8a6. (Accessed 3.6.19)

49 See Mantalero 2018, above, section 1.3 for more detail.

business customers and are often contractual in nature: accordingly, requirements specification often generates long lists of 'shoulds' stated in a natural language. In contrast, where consumers are targeted end-users, their work, play and social needs are typically the focus of requirements analysis. Smaller tech companies, particularly start-ups, identify requirements analysis as the major component of SE concern[50] and are heavily dependent on agile (or even craft) based approaches to development that have weaker contractual type requirements analysis and weaker audit trails. However, it has at least been realised that affected stakeholders, *beyond* those of the customer and end-user, should be identified and involved as participants in requirements analysis and design, giving rise to participatory design[51]. It is also the case that professionals involved in requirements analysis have a diversity of backgrounds; not only computer science and engineering but also psychology, sociology, and other social sciences, and therefore might be extended to include those with legal training. In order to meet Principle 1, software engineering practice of AI systems must meet the following challenges:

a. How to consider human rights requirements for all stakeholders, not only users or customers, but also as individual rights-bearers entitled to equal concern and respect, in auditable requirements description and requirements specification documents?

b. How to train and and employ design professionals who can bring human rights centred design methods to system design and requirements specification?

4.2 Understanding, collecting and analysing data

The processes for acquiring, selecting and modelling data that are required by AI Systems create their own human rights challenges. These challenges require attention at several levels, including the way in which problems are framed during requirements analysis. A human rights-centred approach to design that meets Principle 1 must take due account of human rights risks when building AI System requirements. For example, many commercial AI Systems are designed to utilise data-driven 'hypernudges' to channel user attention and action in directions beneficial to the system owner.[52] These potentially threaten individual autonomy, dignity and the right to liberty and freedom of thought yet these human rights risks are not currently taken into account in requirements analysis processes. **Bias in data sampling, modelling and attribute selection is another major problem.** The use of machine learning (ML) techniques generate many opportunities for bias and discrimination to inadvertently affect the outputs they produced which may threaten the right to the equal protection (the right to non-discrimination).[53] These include biases of the algorithms'

---

[50] Klotins, EU, and Gorschek, T. (2019) 'Software Engineering Antipatterns in start-ups.; *IEEE Software* 36(2): 118-126.

[51] Simonsen, J., and Robertson, T. (Eds.). (2012). *Routledge International Handbook of Participatory Design*. Routledge.

[52] Yeung, K. (2017). 'Hypernudge': Big Data as a mode of regulation by design. Information, Communication & Society, 20(1), 118-136.

[53] Protocol No 12 ECHR Article 1 provides that 'the enjoyment of any right set forth by law shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.' See also Art 21 CFEU.

developers, bias built into the model upon which the systems are generated, biases inherent in the data sets used to train the models, or biases introduced when such systems are implemented in real-world settings.[54] In response to these concerns, a growing body of work concerned with devising technical approaches for countering such bias has emerged, but this has yet to move out of the lab into software development settings.

4.3 Verification

Verification concerns processes for checking whether software meet the specified requirements. In other words, does the software satisfy the output of the requirements analysis? Verification can involve the use of formal methods (logic) to check that software, or a model of software, does not contain errors. Formally verified software performs the required functions and nothing else for all possible inputs with verifiable evidence. Verification is used, for example, in the aviation industry and to some extent in other safety critical systems. Verification has also seen some successes in more agile software development environments.[55] It does not guarantee that, for example, a plane will not crash, but can guard against undesirable conditions occurring by virtue of misconceived models or poorly written software. The application of verification is mandated by certification authorities in some sectors (e.g. aviation) but not in others. It is also used in some sectors (e.g. ship design) because the commercial costs of error are relatively high. Where it is mandated, processes are typically subject to audit by a government authority (e.g. the Federal Aviation Administration (FAA) or the European Aviation Safety Agency (EASA)). Verification is particularly relevant to Principles 1 (design and deliberation) and 2 (assessment, testing and evaluation). We ask two questions: (1) Can AI Systems be formally verified? (2) Can verification be human rights centred? The answer to the first question is negative, at least with current methods[56], although this is an active area of research. Note that AI is not used in safety critical systems precisely because the software cannot be verified. AI Systems are difficult to verify for several reasons. In particular, programs automatically generated by machine learning are coded in different form to hand-coded computer programs and which existing verification methods are not designed to work with. Relatedly, these programmes are typically more complex and have a much higher level of dimensionality in comparison with hand-coded programmes, so that verification approaches may not be computationally tractable. In addition, machine learning can be used in deployed systems to adapt their behaviour in real-time, so that effective verification would need to be continually repeated in the use context. It is therefore impossible for current verification techniques to be human rights centred. Nevertheless, human rights centred verification may still play a role in the design of validation procedures (more on this below). Human rights-centred verification of AI systems is likely to require many years of research before it influences practice. This research should be designed to answer the following challenges:

---

[54] Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*. https://doi.org/10.1177/2053951717743530.

[55] Calcagno, C, Distefano, D, Dubreil, J, Gabi, D, Hooimeijer, P, Luca, M, O'Hearn, P, Papakonstantinou, I, Purbrick, J and Rodriguez, D. (2015) 'Moving fast with software verification.' In *NASA Formal Methods Symposium*,. Springer, Cham. 3 -11

[56] Russell, S, Dewey, D and Tegmark, M (2015) 'Research priorities for robust and beneficial artificial intelligence.' *Ai Magazine* 36(4): 105-114.

c. What are the limits of verification with respect to AI systems and requirements concerning human rights beyond safety?

d. How can formal methods be used to verify AI Systems substrate, e.g. the operating systems and learning software,[57] at least to ensure that AI systems are acquired as intended?

e. How can formal methods be used to advance AI Systems testing (more below)?

4.4 Cybersecurity by design

Cybersecurity concerns the protection of computer systems, data and cyber-physical systems from intrusion, theft, or damage. Cybersecurity by design focuses on the need for security from software foundations and therefore for security considerations in the requirements analysis. Further, some have advocated the use of, for example, formal verification methods early in the design process.[58] Unfortunately, cybersecurity by design has not been a consideration in software engineering until relatively recently and many deployed systems and practices suffer as a consequence. However, this is changing, often in response to legislation, and early efforts at building a regulatory and engineering infrastructure may provide a way forward for human rights-centred design. Inevitably cybersecurity by design for AI Systems faces the same challenges as verification discussed above. Further challenges are documented in a recent NSF report, although that report focuses on privacy rather than on human rights in general.

4.5 Validation

Validation methods are used to assess whether the behaviour of a software system meets stakeholders' needs. Validation is not simply about checking the behaviour of the system against the written specification, it is very much a rigorous empirical process that must generate data relevant to understanding whether a system is fit for purpose. Two specific forms of validation that are conducted extensively in the software industry are:

(a) Penetration testing.

Penetration testing is a commissioned cyber-attack, conducted by an internal or external agency. It is one method used to ensure the security of software systems and data. Penetration testing is sometimes automated and there are standard tests that are legally mandated in some industries (e.g. the Payment Card Industry). A particular focus for penetration testing is privacy validation. However, the value of privacy is sometimes regarded as an absolute value, to be protected at all costs, and which may not reflect the rights-balancing approach enshrined in the way in which the right to privacy is understood within a human rights approach. Penetration testing has not, to our knowledge, been applied to AI systems but recent demonstrations of how AI systems can be spoofed with adversarial

---

[57] Russell, S, Dewey, D and Tegmark, M (2015) 'Research priorities for robust and beneficial artificial intelligence.' *Ai Magazine* 36(4): 105-114.

[58] Chong, S., Guttman, J., Datta, A., Myers, A., Pierce, B., Schaumont, P., ... & Zeldovich, N. (2016). *Report on the NSF workshop on formal methods for security*. *arXiv preprint arXiv:1608.00678*.

attacks[59] suggesting that these systems will come with new, unanticipated, vulnerabilities. In order to meet the needs of Principle 2, penetration testing must meet the following challenges:

(i) Acknowledge and explicitly address trade-offs. Current practice and literature emphasises privacy in ways that may disproportionately threaten the protection of other rights or legitimate collective interests when they come into conflict in specific contexts and circumstances;

(ii) Address the problem of how to counter the threat of Adversarial AI (also known as Offensive AI), particularly as it is likely to be applied to AI systems.[60]

(b) User Experience Design (UX)

User experience designers often play key roles in requirements analysis and also in empirical validation. Tasks and systems (artefacts) often co-evolve and UX designers provide critical feedback on the effectiveness of existing designs as well as ideas for future designs. Typically, they focus on ensuring that system use is useful, pleasurable, rewarding, and efficient. UX designers are also tasked with seeing things from the user's perspective rather than from the service provider's perspective. Methods include the use of scenarios and personas which provide means to curate stories about context of use and potential users. Participatory design has grown in importance, providing one way in which human values from outside the industry can influence design. Human-centred design research has had a strong influence on UX design practice, but methods are not currently configured to embrace human rights concerns in the form recognised under international human rights law. Much has been made of the need for UX designers to consider social and physical context of use[61] and there are a number of ethical and value motivated influences[62], including against bias[63], from feminism[64], from accessibility and diversity research. But, little work is done on the democratic context of use and therefore on human rights from a legal and constitutional perspective. Accordingly, key questions include: What is the future of UX design for AI Systems? How can UX design move beyond its current focus on social and physical contexts to embrace democratic and civic structures (including respect for human rights) as important sources of constraint? Can

---

59 Hutson, Matthew. "Hackers easily fool artificial intelligences." (2018): 215-215.

60 Brundage, M. et al. (2018). The Malicious Use of AI: Forecasting, Prevention and Mitigation.

61 Heath, C and Luff, P (2000). *Technology in action*. Cambridge University Press; Benyon, D, Turner, P and Turner, S (2005) *Designing interactive systems: People, activities, contexts, technologies*. Pearson Education.

62 Friedman, B., and Hendry, DG (2019) *Value sensitive design: Shaping technology with moral imagination*. MIT Press; Friedman, B (ed.) (1997) *Human values and the design of computer technology*. No. 72. Cambridge University Press.

63 Friedman, B. and Nissenbaum, H (1997) 'Software agents and user autonomy.' *Agents* 466-469.

64 Bardzell, S. (2010) 'Feminist HCI: taking stock and outlining an agenda for design.' In *Proceedings of the SIGCHI conference on human factors in computing systems*, 1301-1310. ACM.

methods used in the human rights community for engaging people with human rights thinking[65] contribute to UX approaches to AI Systems design?

4.6 Appropriation

Software systems are not only designed, they are also appropriated by users for unanticipated tasks and unanticipated contexts. This productive aspect of human use of technology may prove particularly problematic for the governance of AI Systems. ADM systems have hitherto been built for one specific social context, but this context is rarely communicated in a robust way that assures that it is only used in this setting. For example, as Zweig and Krafft discuss, the software COMPAS is used in criminal justice systems for pre-trial assessment (such as bail decisions or decisions to prosecute) but was originally built for post-trial assessment.[66] Further work is needed to investigate how appropriation of AI Systems can be appropriately governed (Principle 3).

4.7 Algorithmic transparency and inspection

At present there is no general requirement for algorithm inspection but recent cases suggest that a systematic approach should be considered in order to provide assurance of the trustworthiness of algorithmic systems, particularly those which directly and adversely affect the rights of individuals. In the context of ADM systems in the US criminal justice system, this has been especially problematic where the algorithms have been developed by commercial software providers claiming intellectual property protection, enabling them to assert rights of confidentiality and secrecy over their algorithms.[67] In contrast, the development and implementation of the HART algorithm used in the UK by Durham police force to make custody decision has been much more open, it has not, however, been made available for public scrutiny.[68] While there are legitimate concerns about 'gaming' that may justify refraining from full public disclosure of certain algorithms, at least in high risk contexts where human rights are seriously threatened, regulators must have legal powers to inspect algorithms and datasets (supporting Principle 3).[69]

4.8 Instrumentation and logging

---

65 See news article 'Good Human Rights Stories Coalition Launched' by European Union: External Action, September 2018.

66 Zweig, KA, Wenzelburger, G and Krafft, TD (2018) 'On Chances and Risks of Security Related Algorithmic Decision-Making Systems. *European Journal for Security Research* 3: 181-203.

67 See State of Wisconsin v Loomis (2016) 881 N.W.2d 749 (Supreme Court of Wisconsin).

68 Durham Police have been reported that they 'would be prepared to reveal the HART algorithm and the associated personal data and custody event datasets to an algorithmic regulator' (*Wired,* 1 .3.18), cited by Council of Europe PACE AS/Jur (2019) 20 'Justice by Algorithm – the role of AI in policing and criminal justice'.

69 But see the Canadian Directive on Automated Decision-Making 2019 which applies to the Canadian federal government's use of automated decision-making systems including machine learning and predictive analytics which, among other things, imposes requirements for the release of any custom source code that is owned by the Government of Canada.

Bryson and Theodorou argue that logging should be mandated in all 'socially critical' fields.[70] We assert that logging must also be mandated in all 'human rights critical' fields (i.e. all systems identified that pose a 'high risk' of unjustifiably interfering with human rights, particularly when they can do so at scale). Firms are unlikely to keep an audit trail that evidences their problematic actions unless they are legally required to do so (eg mandatory black box recorder requirements in the aviation industry). The importance of maintaining audit trails of system behaviour is essential for maintaining meaningful human accountability over AI systems: human rights-centred design demands instrumentation in systems so that they automatically record and reproduce historical decision-making processes and outcomes: we should mandate this, at minimum, for all safety critical and human-rights critical systems (in support of Principle 4.)

5. Conclusion

This Chapter has highlighted various deficiencies inherent in the prevailing model of voluntary self-regulation for securing 'ethical AI'. It has enabled a 'Pick Your Own' approach to the identification of ethical standards for AI systems, so that there is no clear, agreed set of ethical standards within the tech industry. This has resulted in conceptual incoherence, particularly because the norms identified in any given 'ethics code' have not typically been rooted in any explicit vision of the kind of political community which those norms are intended to nurture and maintain. Nor do these ethical codes acknowledge the inescapable tensions and conflict that can arise between ethical norms in specific circumstances, let alone offer concrete guidance concerning how those conflicts should be addressed and resolved, leaving industry unilaterally to resolve (or indeed ignore) as they see fit. The prevailing self-regulatory approach also fails to recognise any need, nor obligation, to seek meaningful input from affected stakeholders or the public at large in identifying the relevant ethical standards or how they should be implemented in the design and operation of AI systems. Finally, these codes lack of any effective governance framework, resources or institutions to independently assess and enforce the relevant ethical standards, let alone ensure redress for those adversely affected and/or sanctions in the event of violation. Accordingly, the prevailing approach to AI ethics amounts to little more than a marketing exercise aimed at demonstrating that the tech industry 'takes ethics seriously' in order to stave offer external regulation. In short, it has failed to deliver 'ethical AI'.

We have argued that an alternative approach to the ethical governance of AI is needed – one that is systematic, coherent and comprehensive, centred on human-rights norms and explicitly grounded in the critical importance of protecting and maintaining the socio-technical foundations required to preserve and nurture our societies as constitutional democratic political orders, anchored in an enduring and inviolable commitment to respect human dignity and individual freedom. We have outlined an approach we call 'human-rights centred design, deliberation and oversight', which we believe has the potential to ensure that, in practice, AI systems will be designed, developed and deployed in ways that provide *genuinely* ethical AI. It requires that human rights norms are systematically considered at every stage of system design, development and implementation (making interventions where this is identified as necessary), drawing upon and adapting technical methods and techniques for safe software and system design, verification, testing and auditing in order to ensure compliance with human rights norms, together with social and organisational approaches to

[70] Bryson, JJ and Theodorou, A (2019) 'How Society Can Maintain Human-Centric Artificial Intelligence.' In Toivonen-Noro, M and Saari E (eds.). *Human-Centered Digitalization and Services,* Springer.

effective and legitimate regulatory governance. The regime must be mandated by law, and relies critically on external oversight by independent, competent and properly resourced regulatory authorities with appropriate powers of investigation and enforcement, requiring input from both technical and human rights experts, on the one hand, and meaningful input and deliberation from affected stakeholders and the general public on the other. This approach draws upon variety of methods and techniques varying widely in their disciplinary foundations which, suitably adapted and refined to secure conformity with human rights norms, could be drawn together in an integrated manner to form the foundations of a comprehensive design and governance regime. Its foundational ethical standards are comprised of contemporary human rights norms, designed around four principles, namely (a) design and deliberation (b) assessment, testing and evaluation (c) independent oversight, investigation and sanction, and (d) traceability, evidence and proof.

This approach will not, however, ensure the protection of all ethical values adversely implicated by AI, given that human rights norms do not comprehensively cover all values of societal concern. In addition, a great deal more work needs to be done to develop techniques and methodologies that are both robust, reliable yet practically implementable across a wide and diverse range of organisations involved in developing, building and operating AI systems, and which work effectively to ensure that compliance with human rights norms is evaluated and operationalized at each stage of system design, development and deployment. There are also very considerable challenges in establishing an overarching legal and institutional governance framework that will ensure that AI systems (particularly those appropriately regarding as posing substantial threats and risks to human rights) can be subjected to meaningful and effective scrutiny by competent and independent regulatory authorities endowed with suitable powers of investigation and sanction, and to develop a systematic approach for integrating these different methodologies and requirements into a unified governance framework which enables meaningful public input and deliberation by affected stakeholders in the design, development and implementation of AI systems.

We hope these challenges will not prove insurmountable. Yet their magnitude should not be underestimated, and solving them will require sustained and systematic research and investigation over a long-term time horizon. Our proposal springs from the premise that it is theoretically possible to translate human rights norms into software design processes and into software requirements that can adequately capture the functionality and constraints that give effect to what are often highly abstract human rights norms. We suspect that some rights will be more readily translatable into software and system requirements, such as some rights to due process (such as rights to contestation and rights to an unbiased tribunal particularly when AI systems have been used to inform or automate decisions about individuals), the right to privacy and rights to freedom from unlawful discrimination, while others are likely to be fiendishly difficult to 'hard-wire', such as the right to freedom of expression, freedom of conscience and freedom of association. Because human rights are often highly abstract in nature and lacking sharply delineated boundaries given their capacity to adapt and evolve in response to their dynamic socio-technical context, there may well be only so much that software and system design and implementation techniques can achieve in attempting to transpose human rights norms and commitments into the structure and operation of AI systems in real world settings[71].

[71] Human rights-centred design, deliberation and oversight should *not* be confused with attempts to design computational systems so that they design-out the possibility of non-compliance with the law, which entails translating human rights concepts into formalizable mathematical concepts that can be hard-coded into computational decision-making systems (and which Hildebrandt

Our approach necessitates research and cooperation in AI design, development and implementation between computational, engineering and technical specialists and legal experts with considerable competence and fluency in human rights discourse and jurisprudence. It means, in effect, that those tech designers, developers and engineers involved in building AI systems acquire a deeper understanding of human rights commitments, and the underlying constitutional framework in which they are embedded, in order to identify how to undertake system design, testing, and implementation in ways that are consistent with our democratic constitutional architecture[72]. At the same time, human rights experts will need to acquire sufficient technical competence in the design, architecture, development and implementation of AI systems both in theory and in real-world practice in order to discharge the advisory and assessment duties that we anticipate will be required at every stage of the AI product lifecycle. Yet most contemporary university programmes in law and in computer science and data science currently lack serious and sustained interdisciplinary training. Even if researchers do succeed in developing the requisite techniques, methodologies and organisational and institutional governance frameworks that are capable of forming the foundational elements of human-rights centred design, deliberation and oversight, a cadre of professionals with the requisite expertise and training will also be needed to work with the tech industry in order to implement them into real world practice. Accordingly, our universities must create, nurture and deliver sustained interdisciplinary training and education to undertake the kind rigorous, creative and problem-oriented interdisciplinary research and co-operation that our approach will require, and to equip professionals with the skills, capacities and commitment to embed the core principles of our approach into the AI systems which will increasingly configure and mediate countless dimensions of our everyday human experience. Although AI began decades ago as an interdisciplinary field, it has since become a technical discipline. Yet given the increasing and rapidly expanding application of powerful AI systems in and across many social domains with

---

refers to as 'Legal by Design'). Rather, we anticipate that our vision of 'Human rights-centred design, oversight and deliberation' will incorporate what Hildebrandt refers to as 'Legal Protection by Design' (LPbD) by developing techniques, methods and governance frameworks that can ensure that computational systems make available to individuals a suite of capacities, rights and meaningful opportunities necessary to provide the kind of substantive legal protection currently offered by the contemporary rule of law. According to Hildebrandt, LPbD seeks to ensure that legal protection is not "ruled out by the affordances of the technological environment that determines whether or not we enjoy the substance of fundamental rights", emphasising the need for democratic participation in the design and operation of complex socio-technical systems that configure our everyday environments, and that those subject to such LPbD should be able to contest its application in a court of law, and hence entails foundational requirements that data-driven decisions affecting individuals should be transparent, justified and contestable. In so doing, "LPbD seeks to ensure that the practical capacity for individuals to exercise their human rights enabled by computational systems reflect the dynamic evolution of human rights norms in order to ensure effective protection as the societal and technological context continues to change over time": M Hildebrandt (2019) *Law for Computer Scientists.* Chapter 10. Available at https://lawforcomputerscientists.pubpub.org (Accessed 18.6.19).

[72] See M Hildebrandt (2019) *Law for Computer Scientists*, Oxford University Press. As Borgesius has observed, we need CS research aimed at investigating how AI systems might be designed so that they respect and promote human rights, fairness and accountability, as well as more normative and legal research: Borgesius, Frederik Zuiderveen *Discrimination, Artificial Intelligence and Algorithmic Decision-Making,* Council of Europe, Directorate General for Democracy, 2018, https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73, accessed June 3, 2019 at 69. See also AI Now Report (2018). Available at https://ainowinstitute.org/AI_Now_2018_Report.pdf at 10

the capacity to operate automatically and at scale, study and research into AI must expand to include the social and humanistic disciplines to equip tech professionals with the expertise and sensitivities required to attend seriously to social contexts and to anticipate and identify potential threats and risks these systems might generate when applied to human populations.

Our approach is also likely to several confront significant cultural challenges before it is actively taken-up. These include serious obstacles to systematic implementation into product development lifecycles for AI. Although there are developed software engineering techniques and practices that rely on mathematical proof that can verify that software systems meet certain specifications and are in use, particularly for safety-critical systems, contemporary software development remains largely a 'craft' activity[73] associated with cultural norms in which creativity and freedom to tinker (or 'hack') are widely shared and enthusiastically celebrated. Having spent his professional lifetime providing expert evidence in legal cases in which very large sums of money have been lost due to failed IT system projects, distinguished software engineer, Martyn Thomas laments the fact that software engineering has yet to mature into a professional engineering discipline, committed to robust technical methods and standards and high levels of professional integrity that characterise the so-called 'noble professions'.[74] Yet if software development remains a predominantly amateur activity that celebrates its capacity to 'move fast and break things', an ethic famously championed by Facebook founder Mark Zuckerberg, then our proposed governance regime is unlikely to take root.

We can readily anticipate objections to our proposal, asserting that as a general, legally mandated regulatory regime, it will stifle innovation and sound the death knell for tech start-ups. Yet there is ample evidence to demonstrate that legal regulation may *foster* rather than stifle socially beneficial tech innovation. For example, the introduction of mandatory environmental laws imposing limits on emissions was an important catalyst in the emergence and development of a competitive market for emission reduction technologies. At the same time, the enactment of the EU's General Data Protection Regulation (GDPR) applies to all personal data collectors and processors: from fledgling start-ups through to the Digital Titans. While it may be too early to tell whether the GDPR has led to a decline in tech start-ups and SME growth, it is worth nothing that there are growing calls in the USA to enact a legal data protection regime that can provide equivalently high levels of protection for US-based data subjects. More importantly, however, if 'ethical AI' is to be anything other than a marketing exercise that echoes the hollow claims associated with 'corporate social responsibility', then wholesale change in the tech industry's cultural attitudes will be required and do much more than pay lip service to human rights. Nor can the obligations of AI system developers and operators discharge their duties arising under our proposed governance regime simply by employing a legal expert willing to certify that, to the best of her knowledge and understanding, the system is compliant with human rights standards. In other words, the role of the human rights expert is not that of the hired gun who formulates arguments to assure regulators that her client's system is legally compliant. Rather, it will be necessary to foster a language and culture of 'human rights consciousness' into the tech industry, so that those involved in the design, development and implementation of AI systems regard human rights

---

[73] Thomas, M. (2015) 'Should We Trust Computers?' *Gresham Lectures*. 20 October. London.

[74] Thomas, M (2018) 'Computers and the Future' *Gresham Lectures*, 12 June. London.

AI

compliance as part of their professional remit, rather than a 'niche' problem to be handed-off to legal experts.

Finally, we locate our proposal as only one important element in the overall socio-political landscape needed to build a future in which AI systems are compatible with liberal democratic political communities in which respect for human rights and the rule of law lie at its bedrock. Both more public debate and global cooperation is required. As Bryson and Theodorou observe:

> "The second special problem of AI is not actually unique to it but rather a characteristic of ICT more generally. ICT, thanks to the internet and other networking systems operate transnationally, and therefore affords the accumulation of great wealth and power, while simultaneously evading the jurisdiction of any particular nation. This means that appropriate regulation of AI requires transnational cooperation. Again, the process to establish transnational agreements, treaties and enforcement mechanisms is nontrivial, but already known and already under way.[75]"

In other words, there is also a need for political will and leadership at the national and transnational level in order to bring about the political, social and technical cooperation and investment that will be needed, given that AI systems have the capacity to operate across national borders without technical difficulties. In short, overcoming the many obstacles to cooperation - at the disciplinary level, the organisational level, the industry level, and the policy-making level will all be needed if we are to bring an end to ethics washing, and deliver on the promise of 'ethical AI' in real world settings.

Bibliography

Algorithm Watch *AI Ethics Guidelines Global Inventory* (2019) Available at https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/

Council of Europe (2019) Consultative Committee on the Convention for the Protection of Individuals with regard to Automating Processing of Personal Data (T-PD) *Guidelines on Artificial Intelligence and Data Protection*, T-PD(2019)01, Directorate General of Human Rights and Rule of Law

Hildebrandt, Mireille (2015) *Smart Technologies and the End(s) of Law*. Edward Elgar, Cheltenham.

Hopkins, Andrew (2012) "Explaining Safety Case". Regulatory Institutions Network Working Paper 87. Available via SSRN network.

Kloza, Dariusz, Niels van Dijk, Raphaël Gellert, István Böröcz, Alessia Tanas, Eugenio Mantovani, and Paul Quinn. "Data protection impact assessments in the European Union: complementing the new legal framework towards a more robust protection of individuals." (2017).

Mantalero, Alessandro (2018) *AI and Data Protection, Challenges and possible remedies.* Study for Council of Europe, (2018), https://rm.coe.int/artificial-intelligence-and-data-protection-challenges-and-possible-re/168091f8a6, accessed June 3, 2019

Nemitz, Paul. "Constitutional democracy and technology in the age of artificial intelligence." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, no. 2133 (2018): 20180089.

[75] Bryson, Joanna J., and Andreas Theodorou. "How Society Can Maintain Human-Centric Artificial Intelligence." *Human-Centered Digitalization and Services, M. Toivonen-Noro, and E. Saari (eds.). Springer* (2019) 16-17.

Raso, Filippo A., Hannah Hilligoss, Vivek Krishnamurthy, Christopher Bavitz, and Levin Kim. "Artificial Intelligence & Human Rights: Opportunities & Risks." Berkman Klein Center Research Publication 2018-6 (2018).

Rieke, A, Bogen, M and Robsinson, DG, *Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods*, An Upturn and Omidyar Network Report, 2018, https://www.omidyar.com/insights/public-scrutiny-automated-decisions-early-lessons-and-emerging-methods, accessed June 3, 2019

Yeung, Karen "A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework" Council of Europe MSI-AUT committee study draft (2019) Available at https://rm.coe.int/a-study-of-the-implications-of-advanced-digital-technologies-including/168094ad40

Bibliography 260 words
Total = 10,618 words (excl footnotes and bibliography).