International Human Rights, Artificial Intelligence, and the Challenge for the Pondering State: Time to Regulate?

José-Miguel Bello y Villarino & Ramona Vijeyarasa[1,2]

Keywords: Human Rights, Artificial Intelligence, Women Rights, Automated Decision Making, Decision Support Systems, Monitoring, Regulation, GBV, Nadia, VioGén, Deepfakes

This article has been published in the *Nordic Journal of Human Rights* (May 2022) https://doi-org.ezproxy.library.sydney.edu.au/10.1080/18918131.2022.2069919

Abstract

This article looks at the risks and advantages of early regulation of artificial intelligence (AI) from an international human rights (IHR) angle. By exploring arguments from scholarly and policy papers from various jurisdictions on possible approaches to the challenges posed by AI, the authors identify a current trend among states to wait rather than proactively regulate. Acknowledging the few notable exceptions, such as a recently proposed EU regulation, the authors draw on well-established international human rights principles to challenge the idea that there is a reasonable or legitimate case to 'wait and see'. In order to outline the IHR implications states will have to grapple with in the years to come as AI grows in usage, the article presents three examples of AI systems which provide a women's rights lens on the issue. It argues that the absence of adequate regulations in the AI domain may itself be a violation of international human rights norms, reflecting a state of play where governments have relinquished their obligations to protect, fulfil, and remedy. However, given the limited likelihood that regulatory actions will occur in the short term, it concludes by proposing an alternative intermediate step. Starting immediately, the IHR monitoring framework should demand that states systematically assess and report on their readiness to deal with human rights risks derived from the deployment of AI systems. This will limit such risks and, potentially, result in a more precise identification of the longer-term needs of regulation.

1. Introduction

In early 2020, international human rights (IHR) experts were invited to participate in a Glion Human Rights Dialogue on the issue of human rights in the digital age. The moderator challenged them to discuss the 'relative merits of regulatory versus self-regulatory (business-led) approaches to mitigating the risks that digital technology can

[1] This work was partially supported by the Australian Research Council under Grant CE200100005 and a Fulbright-Schuman fellowship (Bello y Villarino) and the Women's Leadership Institute of Australia Fellowship under Grant PRO20-10998 (Vijeyarasa). The authors express their gratitude to all funders who enabled this research and to the anonymous reviewers.

[2] The authors declare no substantive conflict of interests.

pose to democracy and civil and political rights'.[3] The discussion was conducted under the Chatham House Rule, but the final report noted a wide agreement on the principle that 'States are responsible for setting the broad rules' and a 'widely held view' that 'governments are not fulfilling this responsibility, either individually or collectively, including at the UN'.[4]

Given this conclusion, it is surprising that among the many pieces of recent academic and policy writing on the interaction between human rights and AI,[5,6] none has openly discussed the question of whether the existing IHR framework actually demands action from states. That is, whether there is an IHR requirement that states must, at the very least, set the 'broad rules' to ensure that the use of AI by private and public authorities will be adequately addressed due to the risk it poses to human rights. This is a very different question from that commonly considered in the literature about how the use of AI has affected and will affect human rights. We will not cover many of the issues often canvassed in those papers that delve into the human rights implications of the digital world. We will spare readers any discussion of freedom of speech and democracy, considering beyond the scope of the article any observations about Facebook, Twitter, or former US President Donald Trump as, we believe, this is not representative of the challenge that AI supposes for the IHR framework.

[3] Glion Human Rights Dialogue, 'Human Rights in the Digital Age: Making Digital Technology Work for Human Rights' (Universal Rights Group 2020) 21 <https://www.universal-rights.org/urg-policy-reports/human-rights-in-the-digital-age-making-digital-technology-work-for-human-rights/>.

[4] ibid.

[5] European Union Agency for Fundamental Rights, '#BigData: Discrimination in Data-Supported Decision Making' (2018) <https://fra.europa.eu/en/publication/2018/bigdata-discrimination-data-supported-decision-making> accessed 30 July 2021; Filippo A Raso and others, 'Artificial Intelligence & Human Rights: Opportunities & Risks' (Social Science Research Network 2018) SSRN Scholarly Paper ID 3259344 <https://papers.ssrn.com/abstract=3259344> accessed 23 September 2021; Eileen Donahoe and Megan MacDuffee Metzger, 'Artificial Intelligence and Human Rights' (2019) 30 Journal of Democracy 115; Steven Livingston and Mathias Risse, 'The Future Impact of Artificial Intelligence on Humans and Human Rights' (2019) 33 Ethics & International Affairs 141; Mathias Risse, 'Human Rights and Artificial Intelligence: An Urgently Needed Agenda' (2019) 41 Human Rights Quarterly 1; Rowena Rodrigues, Konrad Siemaszko and Zuzanna Warso, 'SIENNA D4.2: Analysis of the Legal and Human Rights Requirements for AI and Robotics in and Outside the EU' (Zenodo 2019) <https://zenodo.org/record/4066812> accessed 12 August 2021; A Renda, 'Europe: Toward a Policy Framework for Trustworthy AI', *The Oxford Handbook of Ethics of AI* (2020); Karen Yeung, Andrew Howes and Ganna Pogrebna, 'AI Governance by Human Rights–Centered Design, Deliberation, and Oversight', *The Oxford Handbook of Ethics of AI* (Oxford University Press 2020); OHCHR, 'A Human-Rights-Based Approach to Data' <https://www.ohchr.org/Documents/Issues/HRIndicators/GuidanceNoteonApproachtoData.pdf> accessed 19 July 2018; OHCHR, 'Artificial Intelligence Ensuring Human Rights at the Heart of the Sustainable Development Goals' (10 March 2021) <https://www.ohchr.org/EN/NewsEvents/Pages/ArtificialIntelligence-SDGs.aspx> accessed 12 July 2021, just to mention a few.

[6] Interestingly, we could not find any examples regarding this debate in official documents in China. Note their absence, for example in State Council of China, 'China's New Generation of Artificial Intelligence Development Plan (Non-Official Translation)' (30 July 2017) <https://flia.org/notice-state-council-issuing-new-generation-artificial-intelligence-development-plan/> accessed 7 October 2021; Jinghan Zeng, 'Artificial Intelligence and China's Authoritarian Governance' (2020) 96 International Affairs 1441..

e suggest that the existing scholarship tends to miss the bigger question. Some very influential scholars, such as Donahoe and Metzger, suggest that 'perhaps the darkest concerns [for HR] relate to misuse of AI by authoritarian regimes'.[7] We would argue that these probably are the darkest, but not the greatest. Our concerns lie with the everyday normalization of AI, because this usage will affect the widest number of people across the largest number of countries. Moreover, its implications are likely to be felt differently by different groups of rightsholders. Daily usage of AI and the daily risks involved have evident and not so evident implications for human rights, which demands the type of higher-principle inquiry we undertake here: how is AI different as a challenge to human rights, what are the risks, and is it reasonable to assume that regulation is a way (the right way, perhaps) to address them? We grapple with a question largely unaddressed to date, namely whether the absence of regulation is a violation of IHR in and of itself.

This article deals with the issue in a very practical manner. In the following section, we introduce to readers the ways in which AI and HR intersect, setting out the two main systems for which HR is relevant: automated decision-making systems and decision support systems. Having set out the challenges AI poses, in Section 3 we discuss how others have approached HR in relation to AI and the difference in our approach.

In Section 4, we apply a women's rights lens to illustrate some of the regulatory challenges that AI poses for states. AI remains largely under-analysed from the perspective of women's rights, and yet there are specific gendered implications in how AI operates and impacts daily lives. Women's rights provide a useful lens of analysis, but also open up the possibility of extracting lessons for other groups of rightsholders. Our examples demonstrate, for instance, the at times competing nature of the benefits involved for some rightsholders and challenges posed to others. They hint at the limits of law in accommodating intersectional identities when regulating what AI can and should not do. We also pose an important question: how much risk is too much?

Much of our critique stems from the general silence of the IHR system when it comes to AI. At the end of Section 4, we lay out a series of questions that can be considered by states to assess risk, determine future practice, and identify a need for regulation. The 'obligation' to regulate is set out in Section 5. In our conclusion we call for, at the very least, some degree of global intervention from the UN system to delineate minimum standards and good practice for legislation. This, we believe, would be a useful starting point to address one of the world's greatest challenges over the next forty years.

2. What is 'artificial intelligence' and why does it matter for human rights?

[7] Donahoe and Metzger (n 5) 115.

The concept of 'artificial intelligence' has been discussed at such length in the scholarship that a contribution here would not bring further clarity to the idea.[8] Most publications about AI and human rights in academic and institutional settings open with an attempt to define AI.[9] This endeavour naturally leads to a discussion of the 1956 symposium at Dartmouth College in the US, where the term is said to have been coined in a paper co-authored by John MacCarthy.[10] Scholars then tend to note the lack of a widely accepted definition of AI,[11] but acknowledge that is often used interchangeably with 'machine learning' and 'deep learning'.[12]

For the purpose of this text, we do not refer to any concrete techniques or approaches but rather use AI to mean a broad range of data-related technologies[13] that 'can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with'.[14] This definition, inspired by the text of a recent proposal made in the EU context to regulate AI, which we discuss below, stresses two points: first, machines generate outputs within the objectives set by humans. Second, the outputs can take many forms, but their essential aspect is their potential to influence something outside the machine itself, an 'environment'.[15] Imagine, for example, a machine that learns to play chess. If its capacity constantly grows to the point that it can beat any human, but its learning processes are happening in a virtual world *without* contact with humans or other machines, it is a type of AI that does not fit our definition. However, if the machine plays against a human, or is used by a human to analyse options and aid that human in learning how to play or decide on their next move in a game, it fulfils our definition. Some refer to this type as 'algorithmic decision-making' systems 'that support, pre-empt or substitute for human decisions'.[16]

[8] For all, see Michael Haenlein and Andreas Kaplan, 'A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence' (2019) 61 California Management Review 5.

[9] A notable exception is Risse (n 5).

[10] Donahoe and Metzger (n 5) 114.

[11] Raso and others (n 5) 10; Philip Jansen and others, 'SIENNA D4.1: State-of-the-Art Review: Artificial Intelligence and Robotics' (Zenodo 2019) 16 <https://zenodo.org/record/4066571> accessed 12 August 2021.

[12] Livingston and Risse (n 5) 142.

[13] In a report from the relevant EU agency they prefer the term "Big Data", see European Union Agency for Fundamental Rights (n 5)..

[14] European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts 2021, article 3(1). COM/2021/206 final, (21 April 2021).

[15] A similar idea is offered by the Expert Group on Artificial Intelligence at the OECD, 'Scoping the OECD AI Principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD' (OECD 2019) 291 7 <https://read.oecd-ilibrary.org/science-and-technology/scoping-the-oecd-ai-principles_d62f618a-en> accessed 30 September 2021.

[16] Tobias D Krafft, Katharina A Zweig and Pascal D König, 'How to Regulate Algorithmic Decision-Making: A Framework of Regulatory Requirements for Different Applications' (2020) n/a Regulation & Governance 1 <http://onlinelibrary.wiley.com/doi/abs/10.1111/rego.12369> accessed 30 September 2021.

What makes 'AI' different for human rights purposes is not the techniques used—although, again, we will briefly discuss this below—but the value (recognition) humans give the outputs because we think they are good ('intelligent'). From a human rights perspective, the precise value and role attributed to these outputs is paramount.

The main addressee of the IHR framework is the state. How we look at the IHR framework determines, to some extent, what obligations states acquire due to the emergence of AI. Nonetheless, and leaving this discussion aside for a moment, it is a consolidated tenet of IHR law—regional and global—that states are not only mandated to abstain from committing acts in violation of human rights, but they must also ensure that the 'essential rights of the persons under their jurisdiction are not harmed'.[17] This is the case regardless of the origin of the harm—public or private—and normally involves an obligation to actively protect the rights of individuals and to create and maintain avenues to remedy violations ('respect, protect and fulfil').

Until recently, that obligation translated into ensuring that individuals were protected from possible human rights violations derived from the actions and inactions of private persons and other public authorities. However, one of the most frequently noted problems in the AI context is that such harm can originate in systems which escape human control to various degrees, or as the OECD Group of experts put it, in systems with 'varying levels of autonomy'.[18] Put plainly, AI in this sense is beyond the logic of the traditional IHR regime, as the state, bring ultimately responsible for human rights protection, would need to exercise its authority over an entity, a system, which may not even be controlled by private or public persons.[19] Behind these autonomous systems is a vision of AI-driven machines as capable of 'intelligent' decisions without human intervention. The level of intelligence necessary to be 'intelligent' may be debatable,[20] but in any case it is based on a 'standard of human intelligence'.[21] That standard is met when systems that act upon the external environment are perceived by humans to take the best possible action in a given situation.[22]

The second aspect of AI that matters for human rights purposes is that the system acts upon the external environment. In very broad strokes, that action—the output, in our definition—can take two forms that correspond to two types of systems: automated

[17] Christian Tomuschat, *Human Rights: Between Idealism and Realism* (OUP Oxford 2014) 146–147.

[18] OECD, *Artificial Intelligence in Society* (OECD 2019) <https://www.oecd-ilibrary.org/science-and-technology/artificial-intelligence-in-society_eedfee77-en> accessed 7 June 2021.

[19] for a detailed account of the problem see Krafft, Zweig and König (n 16) s 3; for an analysis of the problems this creates to regulate it in general see Simon Chesterman, *We, the Robots?: Regulating Artificial Intelligence and the Limits of the Law* (Cambridge University Press 2021) ch 2.

[20] See a discussion of the Turing test in Chesterman (n 19) 114–115.

[21] Jansen and others (n 11) 12.

[22] Based on the definition in Raso and others (n 5) 11.

decision-making[23] and decision support.[24] Automated decision-making systems are those most commonly associated with AI. Their outputs directly influence the environment they interact with. For example, a system that automatically grants or denies refugee status based on the data it can access about countries and applicants fits into the automated decision-making category. Decision support system outputs do not directly influence the environment, but could be very similar in practice. These systems guide or advise human decisions on the basis of non-banal contributions derived from the processing of a dataset. Continuing with the example of the refugee application, a decision support system could guide an official through the steps for reviewing an application and facilitate their decision by providing information about the applicant's country of origin and offering insights from similar applications already processed by receiving countries in comparable circumstances. It might remind a public servant to consider the relevance of intersectional identities in determining a refugee claim and show media clips collected from the internet about the situation of people from similar groups to the applicant in their country of origin. The purpose would be to ensure that the reviewer considers what the decision support system identifies as relevant factors and to provide pertinent background information, so that applications processed by different people are treated in a similar manner. In the end, however, the decision remains in human hands. Some see the latter systems as the real future of AI.[25]

However, this neat differentiation between automated decision-making and decision support systems can easily blur if the decision support system gives recommendations that humans tend to blindly follow. For example, recommendations from the VioGén system—a decision support system discussed below—were accepted by public officials in around 95% of the cases.[26] This poses questions about the boundaries between automated decision-making and decision support systems, and about whether there is real human autonomy if human discretion is almost never exercised in these settings.

What's more, how the data is collected can also make the distinction between automated decision-making and decision support systems problematic. Imagine that the person reviewing the application for refugee status is unsure about the applicant's honesty and suggests a polygraph test,[27] which the applicant accepts. No one would consider a polygraph test an example of AI, but an AI-driven approach is possible—although not

[23] Ari Ezra Waldman, 'Power, Process, and Automated Decision-Making Symposium: Rise of the Machines: Artificial Intelligence, Robotics, and the Reprogramming of Law' (2019) 88 Fordham Law Review 613, 613.
[24] To use the general terminology coined for management in the 1970s. See Bin Fang, 'Decision Support System (DSS)-Form, Development and Future', *2009 First International Workshop on Education Technology and Computer Science* (2009).
[25] Frank Pasquale, *New Laws of Robotics. Defending Human Expertise in the Age of AI* (Belknap Press 2020).
[26] José Luis González Álvarez, Juan José López Ossorio and Marina Muñoz Rivas, *La valoración policial del riesgo de violencia contra la mujer pareja en España – Sistema VioGén* (Ministerio del Interior Gobierno de España 2018) 56.
[27] Leaving aside the issue of their limited reliability. See American Psychological Association, 'The Truth About Lie Detectors (Aka Polygraph Tests)' (*https://www.apa.org*, 5 August 2004) <https://www.apa.org/research/action/polygraph> accessed 30 September 2021.

very realistic at this stage of the technology. The polygraph could be replaced by a software-based system which has learnt, from many hours of video footage of interviews, to recognize fear. Fear is a key element in the legal definition of who qualifies for refugee status.[28] Therefore, if the system's output is considered a valid measure of fear—i.e., the extent to which the applicant possesses the requisite level of fear—it would have a direct impact on the application. Systems with sensors instead of video recordings that could conceivably replace human judgments about levels of fear.[29] Could the judgment of a human interviewer then legitimately dismiss the AI outcome? Although just a decision support system, its output would be one of the most relevant factors in the application.

At this point, readers have a good grasp of how the use of AI can be a challenge for the IHR framework as we know it, and that this challenge is likely to get more significant over the next forty years. What is less clear is the scale of the challenge and why. In the next section we discuss how others have dealt with that question and the difference in our approach.

3. The artificial intelligence challenge for human rights

In the current scholarship it is relatively well accepted that AI is a challenge for human rights. Some scholars raise concerns about that challenge but understand human rights norms only as forming a boundary around what an AI system can acceptably do.[30] This view, largely reflected in the work of scholars not specialized in human rights or law,[31] sees human rights largely as a manifestation of an agreed common (international) standard that AI systems must respect. For many who adopt this line of thought, the key point is a balancing act between the benefits of the system and the risks to human rights. An assessment of these risks as 'high' or 'very high' triggers an obligation that the designers of such systems 'reconsider and redesign the system and/or proposed business model in order to reduce those risks to a form and level regarded as tolerable'.[32]

One of the co-authors of this article has already explored the limitations of this approach of determining what is tolerable in another paper.[33] It may work well when the balance is between a commercial interest (e.g., a bank using AI systems to reduce the cost of processing mortgages and reducing defaults) and possible human rights violations (the discrimination suffered by a group traditionally mistreated by banks and that suffers

[28] The person needs to have a 'well-founded fear', according to article 1(A)(2) Convention relating to the Status of Refugees 1951.

[29] On a related note, training and testing such a system to reliably recognize fear would also have serious HR implications.

[30] Mark Latonero, 'Governing Artificial Intelligence: Upholding Human Rights & Dignity' (2018) <https://apo.org.au/sites/default/files/resource-files/2018-10/apo-nid196716.pdf>.

[31] Yeung, Howes and Pogrebna (n 5).

[32] ibid 89–90.

[33] José-Miguel Bello y Villarino and Henry Fraser, 'Acceptable "Residual Risks" for Fundamental Rights? Understanding the Keystone of Risk-Based AI Regulations' (2021). Conference paper, presented at 'AI: The New Frontier of Business and Human Rights' (on-line, September 2021).

further discrimination with the implementation of the AI systems). It can also be useful if we think of uses of AI by government entities for public-interest purposes that nonetheless put at risk accepted human rights norms (e.g., massive and indiscriminate face recognition systems for the prevention of crime).

However, as a simplified approach it tends to underestimate the possibility of AI-driven systems simultaneously affecting different human rights in opposite directions (e.g., systems designed to protect some rights which undermine others, an example explored in greater depth below in relation to the Nadia model) and the normative elements in any judgment allowing a system to be put into operation when one small group suffers most of the harm while the benefits are higher but dispersed across society. Conversely, this approach tends to overestimate the usefulness of conducting a cost-benefit analysis for the redesign of those systems when the risks cannot be anticipated (because we do not know what they are or because we do not know how likely they are), or they are very unlikely but so big (fat-tailed) that the only option is not to reconsider or redesign, but to never put the system into operation.[34]

At the other end of the spectrum, the scholarship has identified another big challenge. For these authors, the intersection of AI and human rights is not a matter of balancing costs and benefits for humans but rather requires expanding rights protections to intelligent systems. The debate in this case is mainly ethical and linked to AI systems' increasing intelligence and autonomy.[35] For these authors, the problem is that the human rights regime must be extended and applied to the systems themselves when they reach a level of autonomy that involves a moral status of sorts.[36] The challenge is to prepare ourselves for the moment when systems reach a level of autonomy such that if they were human we would feel they should have rights and obligations. As Livingston and Risse have graphically put it, the Universal Declaration of Human Rights would need to be replaced by a 'Universal Declaration of the Rights of Full Ethical Agents'.[37] A more restrictive but concerning approach affects humans modified (enhanced) by AI systems: where does human stop and machine start? In our view, the challenge here is probably to the idea of human rights as 'human'. The debate is similar to that regarding apes[38] or animals more generally.[39] Although we do not deny it is possible that such a day will come, as it has in the field of animal rights, we will not be facing it in the near future.

[34] See a similar argument in Cass R Sunstein, 'Maximin' (2020) 37 Yale Journal on Regulation 940.

[35] Risse (n 5).

[36] Zeyi Miao, 'Investigation on Human Rights Ethics in Artificial Intelligence Researches with Library Literature Analysis Method' (2019) 37 The Electronic Library 914.

[37] Livingston and Risse (n 5) 151.

[38] Steven M Wise, 'A Great Shout: Legal Rights for Great Apes', *Animal Rights* (Routledge 2008).

[39] Peter Singer, 'Morality, Reason, and the Rights of Animals', *Primates and Philosophers: How Morality Evolved* (Princeton University Press 2009) <http://www.degruyter.com/document/doi/10.1515/9781400830336-010/html> accessed 1 October 2021.

In between these two approaches, a significant body of scholarship and policy documents tries to elucidate the concrete challenges that AI can bring to human rights. Authors tend to propose specific methodologies (often human rights impact assessments) to assess how AI systems can affect human rights in certain examples.[40] Some have extreme positions and call for 'a human rights regime against robotics',[41] but most recognize the need for a balanced approach.[42] Many authors acknowledge that while AI systems are a source of risk for human rights, they could also be an opportunity to improve protection of those rights in line with Human Rights Council Resolution 41/11, which recognized that digital technologies can have positive as well as negative implications for economic, social, and cultural rights. An illustration of the most positive perspective is the open-ended summit organized by the International Telecommunications Union, aptly named AI for Good,[43] but the balance between positives and negatives varies.[44] Authors in this group normally focus on the intersections in particular areas, such as criminal law,[45] media,[46] transnational corporations,[47] or rule of law,[48] among others, allowing for more precise discussions and examples. A common denominator in such approaches is an underlying assumption that existing tools—namely human rights impact assessments—can adequately address the challenges posed by these new technologies.

Finally, AI can bring completely new challenges to our IHR framework because it has the potential to create new domains in which human dignity might be affected. It is difficult to imagine capacities that do not presently exist, but it may be that none of the existing substantive human rights norms would apply. A paramount example here is a system that uses AI to recognize feelings or intentions simply based on how we walk—an advanced type of 'gait recognition'[49] from which particular action is recommended. For example, a

[40] Raso and others (n 5); Australian Human Rights Commission, 'Human Rights and Technology Final Report' (2021) <https://tech.humanrights.gov.au/sites/default/files/2021-05/AHRC_RightsTech_2021_Final_Report.pdf>.

[41] Hin-Yan Liu and Karolina Zawieska, 'A New Human Rights Regime to Address Robotics and Artificial Intelligence' [2016] JusLetter IT s 6.

[42] An overview can be found in Sheshadri Chatterjee, Sreenivasulu N.S. and Zahid Hussain, 'Evolution of Artificial Intelligence and Its Impact on Human Rights: From Sociolegal Perspective' (2021) ahead-of-print International Journal of Law and Management <https://doi.org/10.1108/IJLMA-06-2021-0156> accessed 2 October 2021.

[43] 'AI for Good' (*AI for Good*) <https://aiforgood.itu.int/> accessed 2 October 2021.

[44] Australian Human Rights Commission (n 40) 9.

[45] Aleš Završnik, 'Criminal Justice, Artificial Intelligence Systems, and Human Rights' (2020) 20 ERA Forum 567.

[46] IAP Wogu and others, 'Human Rights' Issues and Media/Communication Theories in the Wake of Artificial Intelligence Technologies: The Fate of Electorates in Twenty-First-Century American Politics' in Thangaprakash Sengodan, M Murugappan and Sanjay Misra (eds), *Advances in Electrical and Computer Technologies* (Springer 2020).

[47] Emilie C Schwarz, 'Human vs. Machine: A Framework of Responsibilities and Duties of Transnational Corporations for Respecting Human Rights in the Use of Artificial Intelligence Notes' (2019) 58 Columbia Journal of Transnational Law 232.

[48] Monika Zalnieriute, Lyria Bennett Moses and George Williams, 'The Rule of Law and Automation of Government Decision-Making' (2019) 82 The Modern Law Review 425.

[49] Tanmay Randhavane and others, 'Identifying Emotions from Walking Using Affective and Deep Features' (2020) <http://arxiv.org/abs/1906.11884> accessed 1 October 2021.

message is sent to police to do a body search of a person because the system believes a correlation exists between their particular way of walking and a chance that they are hiding something. Would this be a violation of privacy as understood in current IHR law? Is privacy the right word? AI systems could be used to analyse voice intonation and grant a certain opportunity to those whose voices are likely to trigger better reactions from the audience. People whose timbres are less motivational would be discriminated against, but would not deserve any kind of protection today; if the system were applied on a mass scale in the service sector, a select group of people may find many job opportunities closed to them. Would this group of less adept voices deserve protection as some post-modern type of disability? Is it reasonable to call it a disability,[50] or do we need a new term? None of the 'usual' human rights (privacy, non-discrimination) seem affected, but these hypotheticals create a feeling that human beings' dignity could be at stake if our way of walking or tone of voice can determine our relationship with the authorities or our social and economic opportunities

Despite all these significant questions, our claim here is that the main challenge for states as the primary entities responsible for human rights protection is not substantive but procedural: have states considered whether they are ready to protect against possible rights violations derived from the application of human rights systems and provide remedies in cases of violations? This is a question of first order before any others are addressed. In our view, IHR law provides a clear mandate to states to actively engage in that consideration.

Article 2 of International Covenant on Civil and Political Rights[51] establishes that '[w]here not already provided for by existing legislative or other measures, each state party to the present Covenant undertakes to take the necessary steps […], to adopt such laws or other measures as may be necessary to give effect to the rights recognized in the present Covenant'. In subparagraph (a) the article states that this obligation extends 'to ensure that any person whose rights or freedoms as herein recognized are violated shall have an effective remedy'. In the words of Tomuschat, this means it is not 'enough for governmental authorities to abstain from committing illegal acts, they must also see to it that the essential rights of the persons under their jurisdiction are not harmed by other private persons'.[52] Tomuschat argues that Human Rights Committee jurisprudence has emphasized 'the protective dimension of the core human rights'[53] and that this approach applies to economic, social, and cultural rights as well as civil and political rights.

[50] Not least at the risk of diluting the great gains that have been achieved in terms of human rights for the communities of people living with disability

[51] Adopted and opened for signature, ratification and accession by General Assembly resolution 2200A (XXI) of 16 December 1966 entry into force 23 March 1976, in accordance with Article 49

[52] Tomuschat (n 17) 146–147.

[53] See General Comment 31, Nature of the General Legal Obligation Imposed on States Parties to the Covenant (UN doc CCPR/C/21/Rev 1/Add 13, 26 May 2004).

In the next section we present three examples connected to women's rights to illustrate the type of thorny questions that states will have to consider when assessing if their human rights legal regime are ready to address the many challenges that the existence of AI systems will involve. Women's rights provide a useful lens due to a growing and relatively well-established appreciation of gender bias in AI, and these examples both tease out unexplored issues and bring visibility to the breadth of sometimes competing human rights concerns. Further, it should become readily identifiable to readers how different groups of rightsholders may be affected by AI-driven machines in particular ways, depending on circumstance. We will then argue that addressing those (and other) questions is the only way to correctly assess whether states are ready to meet their IHR obligations in connection to AI.

4. Artificial intelligence and human rights in practice: the questions for the state through the lens of women's rights

The following three examples elucidate the implications of AI for women's rights. The first two cases demonstrate the repercussions of using AI to achieve public policy objectives connected to human rights advancement. The third, an example from the private sector, demonstrates that AI can be used in direct violation of women's rights norms. The first case we offer, Nadia, demonstrates how some of the rights of women as a group may be put at risk through attempts to advance the rights of others, in this case people living with disabilities. In contrast, the VioGén system is designed to protect a particular set of rights of some women—the right to live free of gender-based violence—but risks interfering with the fundamental rights of other individuals. The third example illustrates the use of a private AI technology by corporations and individuals in a way specifically designed to undermine women's rights: creating 'deepfakes' of naked women and making them freely accessible.

Nadia

Nadia is a virtual assistant developed by the Australian Government's National Disabilities Insurance Agency (NDIA).[54] Trials of NADIA were meant to begin in mid-2017, starting with a 12-month period in which the system was intended to learn[55] before its public release. As of September 2021, the project had been either terminated or stalled, likely due to the government's aversion to the risks it was seen to pose.[56] Nadia

[54] It is perhaps surprising how little attention—scholarly or popular—Nadia has drawn, given the assistant is voiced by actor Cate Blanchett. Blanchett's inclusion in the project may have been designed to bring Nadia and Australia notable visibility, yet the choice of a 'celebrity voice' does raise questions about Nadia's evidence-based design, the voice itself being a key component of the AI technology

[55] Christopher Knaus, 'NDIA Denies Cate Blanchett-Voiced "Nadia" Virtual Assistant Is in Doubt' *The Guardian* (21 September 2017) <https://www.theguardian.com/australia-news/2017/sep/22/ndia-denies-cate-blanchett-voiced-nadia-virtual-assistant-is-in-doubt> accessed 2 October 2021.

[56] Exclusive by political editor Andrew Probyn, 'Government's Blanchett-Voiced AI Venture for NDIS Stalls' *ABC News* (21 September 2017) <https://www.abc.net.au/news/2017-09-21/government-stalls-ndis-virtual-assistant-voiced-by-cate-blanchet/8968074> accessed 25 September 2021.

was developed by the New Zealand company FaceMe, whose main business was commercial. FaceMe described the creation of Nadia as an experience of developing an omni-channel digital employee platform.[57] It was meant to use AI to 'help the NDIA communicate with the hundreds of thousands of national disability insurance scheme participants', giving 'spoken or written answers in 32 languages to thousands of NDIS queries' and learning from those interactions.[58]

Its designer, Marie Johnson, argued in its submission to a Parliament Committee that Nadia's origins laid squarely in the UN Convention on the Rights of Persons with Disabilities, namely the Convention's call to promote communication for people living with disabilities, including through 'human-reader and augmentative and alternative modes, means and formats of communication, including accessible information and communication technology',[59] and for states parties to 'receive and impart information and ideas on an equal basis with others'. [60] However, even if we can identify the ways in which Nadia's use might advance the interests expressed in the Convention, it falls short as an example of 'human rights-by-design' technology[61] or 'design for human rights'.[62] Specifically, the choice of a female voice (that of Australian actor Cate Blanchett) for this voice-activated personal assistant (VPA) demonstrates the clear risk of promoting gender stereotypes and undermining progress towards gender equality.

The Office of the High Commissioner for Human Rights (OHCHR) has named gender-stereotyping a 'pervasive human rights violation',[63] although its 2013 report does not grapple with the types of discrimination at risk in VPA. Design choices for VPAs tend to draw on behavioural economics that reinforce assumptions associating the female gender with feelings of assurance, trust, safety, and placidity. Female voices are a choice; in comparison to associations drawn with the use of a male gendered voice, 'she assists rather than directs; she pacifies rather than incites'.[64] While the phenomenon is more complex in the case of Nadia, as its very purpose was to advance the interests of an already vulnerable group, gendered assistant technologies have been described as a form

[57] Roger Smith, 'Nadia Falters: Teetering Technology in the Service of Access to Justice' (*Law, Technology and Access to Justice*, 6 November 2017) <https://law-tech-a2j.org/advice/nadia-falters-teetering-technology-in-the-service-of-access-to-justice/> accessed 2 October 2021.

[58] Knaus (n 55).

[59] Convention on the Rights of Persons with DIsabilities 2006 s Art. 2.

[60] ibid Art. 21. Stephen Easton, 'Nadia: The Curious Case of the Digital Missing Person' [2019] *The Mandarin* <https://www.themandarin.com.au/106473-nadia-the-curious-case-of-the-digital-missing-person/> accessed 2 October 2021.

[61] Jonathon Penney and others, 'Advancing Human Rights-by-Design in the Dual-Use Technology Industry' (2018) 20 Columbia Journal of International Affairs <https://digitalcommons.schulichlaw.dal.ca/scholarly_works/250>.

[62] Evgeni Aizenberg and Jeroen van den Hoven, 'Designing for Human Rights in AI' (2020) 7 Big Data & Society 2053951720949566.

[63] Office of the High Commissioner for Human Rights, 'Gender Stereotyping as a Human Rights Violation' (2013).

[64] Nora Ni Loideain, Rachel Adams and Damian Clifford, 'Gender as Emotive AI and the Case of "Nadia": Regulatory and Ethical Implications' (Social Science Research Network 2021) SSRN Scholarly Paper ID 3858431 6 <https://papers.ssrn.com/abstract=3858431> accessed 6 September 2021.

of 'digitally-gendered servitude', which risk reifying 'negative and harmful stereotypes around the role of women as secondary to men, who exist simply to serve others'.[65] Perhaps unintentionally and indirectly, Nadia, like Alexa and Siri,[66] promotes a limiting and risky stereotype about women's societal roles.

On this point, a December 2021 UNESCO recommendation notes that persons in vulnerable situations can receive assistance from AI systems, but that such interactions should never objectify or undermine human dignity.[67] 'Diversity and inclusiveness', the recommendation explains, 'should be ensured throughout the life cycle of AI systems, consistent with international law, including human rights law'.[68]

The body of scholarship addressing the gendered implications of VPAs is limited and largely produced by small group of scholars. However, this does not mean that the more tangible manifestations cannot be simply illustrated. From a human rights' perspective, a relevant question here is whether Cate Blanchett would be the choice of voice for an AI system giving instructions to operate a nuclear reactor in case of emergency. What does this say about the deeper gender implications of the outputs of AI systems? Nadia's creation exemplifies one of the core human rights challenges in regulating AI, grappled with further below: potential gender-based harms to one category of rightsholders created through the intention to advance the interests of another.

VioGén

For more than ten years, gender-based violence (GBV) teams led by the police in Spain have been assisted by a computer-based system to assess the risk of recidivism by perpetrators of GBV. This approach exists against a backdrop of ongoing research being carried out in Spain to identify the particular traits commonly present among perpetrators of GBV.[69] By contrast, in other countries, such as Australia, attempts to understand traits common to domestic violence offenders are relatively new, although research suggests that it is indeed possible to identify consistent patterns in the traits of offenders that lead to offending and reoffending.[70] While necessarily part of a broader set

[65] ibid.

[66] Rachel Adams and Nóra Ní Loideáin, 'Addressing Indirect Discrimination and Gender Stereotypes in AI Virtual Personal Assistants: The Role of International Human Rights Law' (2019) 8 Cambridge International Law Journal 241.

[67] UNESCO, 'Recommendation on the Ethics of Artificial Intelligence' para 15 <https://unesdoc.unesco.org/ark:/48223/pf0000380455> accessed 7 March 2022.

[68] ibid 19.

[69] Julie Van Hoey and others, 'Profile Changes in Male Partner Abuser After an Intervention Program in Gender-Based Violence' (2021) 65 International Journal of Offender Therapy and Comparative Criminology 1411.

[70] Shann Hulme, Anthony Morgan and Hayley Boxall, 'Domestic Violence Offenders, Prior Offending and Reoffending in Australia' (2019) No. 580 Trends and Issues in Crime and Criminal Justice 22.

of interventions, studies generally acknowledge that identifying the risk of recidivism is an important approach to protecting victims of GBV.[71]

To better assess the risk of re-victimization in gender-based violence, Spain established the VioGén system, which monitors the experiences of victims and proposes possible protection and assistance measures to public authorities. Generally speaking, this mechanism of risk assessment has been deemed a key part of the relative success—at least in terms of a reduction in absolute number of deaths[72] and recidivism[73]—of the Spanish model to fight GBV.[74] It also responds to the idea of using technology to better address women's rights issues.[75]

The system has had three iterations: (i) a human-driven system; (ii) a system supported by traditional statistics; and (iii) an AI-driven system. The version still in place is 'ii', so the VioGén system is not yet applying an AI model; it relies instead on traditional statistical methods to estimate risk by comparing the case at hand and the historical data in the VioGén database.[76] In 2018, Spain's Ministry of Interior contracted a group of data scientists from various Spanish universities to design a system that could improve the accuracy of the predictions.[77] In test conditions, the statistical system outperforms human psychological assessments[78] and the AI system outperforms both, showing a significant improvement in terms of accuracy,[79] perhaps by 10 or 15 percent. In absolute

[71] P Randall Kropp and Stephen D Hart, 'The Spousal Assault Risk Assessment (SARA) Guide: Reliability and Validity in Adult Male Offenders' (2000) 24 Law and Human Behavior 101, 102.

[72] 72 victims in 2004 against an average of 50 victims in the 2016-2020 period 'Mujeres - Delegación Del Gobierno Contra La Violencia de Género' (2021) <https://violenciagenero.igualdad.gob.es/violenciaEnCifras/victimasMortales/fichaMujeres/home.htm> accessed 23 September 2021.

[73] Decrease in 25% of rates of recidivism according to the data given to the press since the implantation of the risk-assessment system.

[74] The system was created by the Organic Law 1/2004 de Medidas de Protección Integral Contra La Violencia de Género (Integral Protective Measures against Gender-Based Violence) Ley Orgánica 1/2004, de 28 de diciembre, de Medidas de Protección Integral contra la Violencia de Género. 2004 (BOE-A-2004-21760).

[75] In 2017, the Committee on the Elimination of All Forms of Discrimination issued a General Recommendation on gender-based violence CEDAW Committee, General recommendation No. 35 on gender-based violence against women, updating general recommendation No. 19 2017 [UN. Doc. No. CEDAW /C/GC/35]., the third and most recent of the Committee's now 38 recommendations to focus on GBV Ramona Vijeyarasa, 'CEDAW's General Recommendation No. 35: A Quarter of a Century of Evolutionary Approaches to Violence against Women' (2020) 19 Journal of Human Rights 153. In this update to its much earlier 1992 recommendation, technology has a place in the Committee's analysis. GBV 'manifests in a continuum of multiple, interrelated and recurring forms, in a range of settings, from private to public, including technology-mediated settings and in the contemporary globalized world it transcends national boundaries' CEDAW Committee General Recommendation No. 35 para 6.

[76] González Álvarez, López Ossorio and Muñoz Rivas (n 26) 89.

[77] Marta Pinedo, 'Matemáticas e inteligencia artificial contra el maltrato machista' *EL PAÍS* (2 September 2021) <https://elpais.com/sociedad/2021-09-02/matematicas-e-inteligencia-artificial-contra-el-maltrato-machista.html> accessed 23 September 2021.

[78] José Manuel Muñoz Vicente and Juan José López-Ossorio, 'Valoración psicológica del riesgo de violencia: alcance y limitaciones para su uso en el contexto forense' (2016) 26 Anuario de Psicologia Juridica 130.

[79] Ángel González-Prieto and others, 'Machine Learning for Risk Assessment in Gender-Based Crime' [2021] arXiv:2106.11847 [cs, stat] <http://arxiv.org/abs/2106.11847> accessed 23 September 2021.

terms, of the 600,000 women in the Spanish VioGén database, 60,000 to 90,000 could have had a better assessment of their level of risk with the new system.[80]

The system presents three evident risks, however. As with any data-driven system, it can be gamed for spurious purposes (adversarial attack) if someone tampers with the information fed into the system .[81] This could affect the level of protection of the victim and/or the rights to a family life and freedom of movement of the alleged perpetrator, clearly articulated in the European Convention on Human Rights. Second, in cases of failure, the responsibility for that failure is not clear. In 2020, the Spanish Audiencia Nacional found that the state had to compensate the family of Stefany González Escarramán, murdered by a former partner against whom she had been denied a restraining order by the Spanish Guardia Civil (militarized police) on the basis of a risk assessment generated by VioGén. A system designed to protect individuals against grave human rights violations had failed. Third, as the new AI system is being tested, we must ask when is a good time to implement an AI technology to protect human rights. Would Ms. González Escarramán have been one of the thousands of cases whose risk of revictimization would have been better assessed if the AI system was already in place?

System Y

Website Y (not identified here to avoid its promotion) is 'eye-catching for its simplicity'[82]: turn any person into a 'porn star' by uploading their photo onto the website, which uses deepfake technology to swap the person's face into an adult video. After users upload a picture of a face, four AI-generated faces allow a test of the results. If accepted by the user, the system then generates pornographic photos or videos through the superimposition of that face onto videos and photos of pre-filmed women (or a small number of men's bodies) in its database.[83] It appears that the technology behind website Y was acquired by the website owners from the party that had developed it.

[80] Pinedo (n 77).

[81] Andrés Boix Palop, 'Los algoritmos son reglamentos: La necesidad de extender las garantías propias de las normas reglamentarias a los programas empleados por la administración para la adopción de decisiones' 264 <https://repositorio.uam.es/handle/10486/692210> accessed 12 September 2021.

[82] Karen Hao, 'A Horrifying New AI App Swaps Women into Porn Videos with a Click' (*MIT Technology Review*, 13 September 2021) <https://www.technologyreview.com/2021/09/13/1035449/ai-deepfake-app-face-swaps-women-into-porn/> accessed 2 October 2021.

[83] ibid.

The legal problems created by 'deepfakes' are well-known,[84] especially in connection to political persons,[85] which has been the object of specific regulation.[86] Their use for pornography, however, currently seems to be the predominant one. Some estimates suggest that more than 19 out of every 20 deepfake videos on the internet in 2019 were pornographic.[87] One particularity of deepfake's use for pornography is its gendered dimension.[88] Pornography-related attacks on non-media-relevant figures are more likely to target women than men, partly because it is an AI tool most often developed with women's bodies in mind, and partly because the ratio of pornographic (still and animated) images involving women rather than men is unbalanced. Evidence of the non-consensual distribution of pornographic deepfakes of men can be found, but the phenomenon more often involves women.[89]

From a criminal point of view, the regulation of deepfake pornography may not be particularly challenging; it could be addressed in the context of the distribution of intimate images without consent. There are several examples of successful prosecutions in different jurisdictions of these crimes, which can easily extend to images that have 'been altered to appear to show a person's private parts, or a person engaged in a private act, in circumstances in which a reasonable person would reasonably expect to be afforded privacy', as the legislation of New South Wales, Australia now establishes,[90] following a wave of advocacy and policy reform in the area of revenge pornography.[91] Yet many of these changes are still being debated elsewhere, because their implementation may not be that simple in all legal systems.[92]

While the CEDAW Committee has been wavering in its view on what pornography means for women—named in its General Recommendation No. 19 of 1992 as the 'commercial

[84] A brief explanation of the technology and its regulatory relevance can be found in Tyrone Kirchengast, 'Deepfakes and Image Manipulation: Criminalisation and Control' (2020) 29 Information & Communications Technology Law 308.

[85] Andrew Ray, 'Disinformation, Deepfakes and Democracies: The Need for Legislative Reform' (2021) 44 UNSW Law Journal 983.

[86] Kari Paul, 'California Makes "Deepfake" Videos Illegal, but Law May Be Hard to Enforce' *The Guardian* (7 October 2019) <https://www.theguardian.com/us-news/2019/oct/07/california-makes-deepfake-videos-illegal-but-law-may-be-hard-to-enforce> accessed 3 October 2021.

[87] Henry Ajder and others, 'The State of Deepfakes: Landscape, Threats, and Impact' (Deeptrace 2019) 1 <https://regmedia.co.uk/2019/10/08/deepfake_report.pdf>.

[88] ibid 2; Travis L Wagner and Ashley Blewer, '"The Word Real Is No Longer Real": Deepfakes, Gender, and the Challenges of AI-Altered Video' (2019) 3 Open Information Science 32.

[89] See a similar reasoning in Amrita Khalid, 'Deepfake Videos Are a Far, Far Bigger Problem for Women' (*Quartz*, 9 October 2019) <https://qz.com/1723476/deepfake-videos-feature-mostly-porn-according-to-new-study-from-deeptrace-labs/> accessed 3 October 2021.

[90] Crimes Act 1900 No 40 - NSW s 91N.

[91] Noelle Martin, *Online Predators Spread Fake Porn of Me. Here's How I Fought Back* (2017) <https://www.ted.com/talks/noelle_martin_online_predators_spread_fake_porn_of_me_here_s_how_i_fought_back> accessed 2 October 2021.

[92] Kirchengast (n 84). As illustrated by wider Australian law which may still be behind the times. John Davidson, 'Australian Law behind the Times on Deepfake Videos' *Australian Financial Review* (8 July 2019) <https://www.afr.com/technology/australian-law-behind-the-times-on-deepfake-videos-20190703-p523pi> accessed 2 October 2021.

exploitation of women as sexual objects',[93] it is not mentioned in General Recommendation No. 35 that updated the earlier recommendation—the human rights issue here is not pornography. Rather it is the non-consensual use of the imagery of individuals, more often than not women, for the economic and personal benefit of others that has both direct and indirect implications for the rights of the women whose images have been exploited. A significant challenge lies in trying to develop a human rights-based approach to these uses of AI that are born as frontal attacks against the human rights of certain groups. Moving away from the more popular idea of 'revenge porn', which does not encapsulate the problem described here because most of the deepfakes on website Y are not motivated by revenge,[94] our concern is that AI can expressly be used by private parties to undermine the rights of a segment of (or all) the population.

The questions

In our view, the three examples evidence the need for, at the very least, an honest assessment of a state's legal regime to address the main risks that AI systems give rise to. We argue that the obligation to protect in this context requires adequate legislation and goes beyond mere implementation of policies. Having AI-ready legislation therefore rises to the level of a state obligation under the IHRL regime. In this subsection we present several questions that need to be considered when assessing states' legal regimes in relation to the problems illustrated by the above three cases:

1. Is the state prepared to balance the different risks posed to different human rights when AI is deployed? AI often relies on optimization, i.e., finding the best way to achieve a given objective (win the chess game, for example, or minimize the risk of exposure to GBV facing women as a group) according to certain constraints. To do this in the human rights context, the state must be ready to expressly prioritize the rights of some individuals or groups over others. Is it ready, and, if so, based on what parameters? Nadia may have been very helpful for people living with disabilities, but evidently risked the perpetuation of particular gendered norms and therefore went against basic tenets expressed in CEDAW. VioGén is a well-intended AI-assisted tool to accelerate the capacity of the Spanish Government to protect a *broader* number of women. Technically speaking, the system only predicts the risk of revictimization—as it is a victim centred approach—but the decisions it helps shape affect the alleged victim (through the frequency of monitoring and contact with the victim by the authorities, for example), the alleged perpetrator (such as recommendations made by the police to a judge to impose restraining orders), and, potentially, other individuals such as children under the custody of either parent. Its

[93] CEDAW Committee, General Recommendation No. 19, Violence against women (Eleventh Session, 1992) 1992 para 12.

[94] Tyrone Kirchengast and Thomas Crofts, 'The Legal and Policy Contexts of "Revenge Porn" Criminalisation: The Need for Multiple Approaches' (2019) 19 Oxford University Commonwealth Law Journal 1, 3–4.

recommendations therefore affect different human rights of different groups of individuals.

A balancing exercise is often required in the drafting of any regulation. A human rights-based regulation of AI may intentionally or unintentionally favour one group of rightsholders over another, and governments need to be prepared to decide who and what to favour, and who or what to sacrifice. Resorting to neutrality—if that exists[95]—is not a feasible option if the purpose of using an AI system is precisely to address a situation of vulnerability of the human rights of one group.

2. How can the state determine when to replace a human-performed system with an AI-based one, in order to advance the interests of some rightsholders, when that decision may affect or undermine the achievement of other rights or carry new risks? This question is connected to the first one, but differs slightly as it refers to the decision taken to use an AI-based system to replace systems already in place but previously performed by humans. In the case of Nadia and the last iteration of VioGén, states need to determine the factors they will consider when moving to an AI system. This may include, but certainly is not limited to, users' readiness, aggregated general welfare, the protection of most discriminated groups, cost savings, or any combination of the above. Having some guidance at a legal level about when to change, or—more realistically—what to consider when making the decision to change from human to machine could avoid many problems.

3. How can states ensure in their regulation that AI systems consider intersectional perspectives when applied at large scales? The protection of the human rights of individuals that are part of several vulnerable groups may not be correctly identified by AI systems. For example, in the case of VioGén, did the system consider that Ms González Escarramán was not only a female victim and an immigrant in Spain but also a person of colour who may have faced distinct challenges in terms of access to housing?[96] This consideration must be placed in the context of the Spanish system, which gives housing assistance to victims, but not necessarily for those with low-risk of revictimization. Factoring in inequalities that stem from race may be an obvious necessity to a human interviewer, but may not be to an AI-driven system if that information is not in the dataset from which it is learning. A legal mandate to consider intersectionality in AI systems could ensure that the human-to-machine transition does not ignore this problem.

Research demonstrates that the risks of AI are exacerbated when applied to vulnerable groups.[97] Scholars who are beginning to explore the potential for EU law to protect against discrimination in AI-based decision-making have highlighted the inadequacies of

[95] Ramona Vijeyarasa, 'In Pursuit of Gender-Responsive Legislation: Transforming Women's Lives through the Law' in Ramona Vijeyarasa (ed), *International Women's Rights Law and Gender Equality: Making the law work for women* (Routledge, Taylor and Francis 2021) 9.

[96] Patricia J Williams, *The Alchemy of Race and Rights* (Harvard University Press 1991) 146–148.

[97] Ni Loideain, Adams and Clifford (n 64).

EU law's 'unidimensional understanding of discrimination' to accommodate the intersectional forms of discrimination that predictive analytics give rise to.[98] A similar dimension of this problem is illustrated by one study about VioGén, utilizing an online satisfaction questionnaire, already piloted in two previous studies, which showed that 80 per cent of women who participated in the study—among the 1,128 valid questionnaires completed—reported feeling very satisfied with the police performance.[99] Yet both this survey data and the very language of risk assessment beg questions about appropriateness. What do we know of, for instance, the experiences of the remaining 20 per cent of women who reported a less than satisfactory response to police performance? What was their lived reality of GBV? Some domestic violence victim advocates may shudder at the use of a machine to decide an individual's level of risk.

4. Is the domestic legal system prepared to identify and respond to human rights violations derived from the use of AI systems which could be considered minor in terms of gravity, but could be severe in societal terms when AI systems are applied on a larger scale? The risk to women's rights presented by each use of Nadia may be limited. Indeed, it is a challenge to quantify it. Some of us may be so accustomed to the virtual assistants streaming into our homes that we would be hard-pressed to believe that they are ingraining a gender stereotype about women's roles. Yet the cumulative effect of voice-activated assistants like Nadia, in their interaction with millions of people, could be long-lasting, for example, by reinforcing stereotypical, limiting, and unequal roles and responsibilities for women in relation to care. Does the state have legal guidance on these issues or is it best left to individual decision makers? If the latter, how do we address the cumulative effect of many individuals making decisions (e.g., if all state-promoted voice assistants in the care economy were assigned female voices)? Should there be legal guidance about diversity in human interactions of machines the same way that there can be gender or minority quotas for public-sector jobs?

5. How can legal systems be flexible and responsive enough to address new types of violations of human rights which only emerge as a result of these new technologies? The type of technology allowing Y website's mass-scale and cheap modification of videos did not exist ten years ago. AI is going to be applied in ways that we cannot anticipate and risks open violation of a range of human rights. In deciding to regulate, states need to demonstrate a preparedness to respond. The alternative is for the state to place its trust in humans to adapt to these new technologies. State regulation in this area should not be considered excessive or extreme. Many states take precautions to avoid the doctoring of photos to obtain passports , licences, or documents to access public services. In summary,

[98] Raphaële Xenidis and Linda Senden, 'EU Non-Discrimination Law in the Era of Artificial Intelligence: Mapping the Challenges of Algorithmic Discrimination' (Social Science Research Network 2019) SSRN Scholarly Paper ID 3529524 <https://papers.ssrn.com/abstract=3529524> accessed 25 September 2021.

[99] José Luis González and María José Garrido, 'Satisfacción de las víctimas de violencia de género con la actuación policial en España. Validación del Sistema VioGen' (2015) 25 Anuario de Psicología Jurídica 29, 30.

should states ignore this problem and just trust users or should they establish legal regimes that can address these challenges and the potential negative consequences?

6. Whose responsibility is it to monitor these systems, and according to what parameters? This challenge is system-specific, but it is not simply a question of a need to monitor the developments of the private sector, given two of our examples were designed by and for the public sector. In the case of Nadia, the system could in fact meet the expectations of the developer and the National Disability Service in terms of improving the lives of people living with disabilities yet still be oblivious to the harms caused to other rightsholders. VioGén could meet the objectives of reducing recidivism, but possibly at the cost of the rights of alleged violators through the recommendation of an increasing number of restraining orders. A human evaluator may deem the consequences of recidivism sufficiently grave to justify this encroachment, but is the state obliged to monitor that activity to protect the rights of potential perpetrators of GBV? If so, and it does seem logical, then question 1 reappears: how can the state modify the parameters to create a balance between rights? As more AI systems are put in place, is the state ready to humanly monitor all of them, systematically and at regular intervals, for changes in the ways in which they function and develop as they continue to learn by themselves?

7. Does the state need to wait for harm and then offer a remedy, or is it necessary to create rules to guide ex-ante prohibitions of uses of AI as they are highly likely to entail violations of human rights? Among our examples, there is an evident violation of human rights norms in the development and sale of the technology behind website Y. We argue that the severe implications for a range of human rights in that case creates an evident burden on the state to step in. A key question to consider is whether any clear existing rules can inform the decisions of private actors about how to guarantee respect for human rights in the use of the technology being created and how much graver the risk is when an AI-based technology replaces what can be done by humans. Photoshopping images was a labour-intensive task a few years ago, but now it can be done on a massive scale with limited human intervention. The speed at which the technology has advanced—with most phones now sold enabling individuals who have few technical skills to alter photos with their fingertips—may demand a regulatory approach that errs on the side of caution. In other words, if regulation cannot keep up with the pace of technological developments, a human rights-based mandate can minimize risk without gravely hindering scientific progress.

8. Are states ready to cooperate to effectively address uses of AI designed to undermine human rights which cannot be regulated or policed domestically? In cases like website Y, single states will necessarily be limited in their ability to prevent human rights violations.

At some near point, the type of international cooperation that we see, for example, in tackling child pornography[100] will need to be considered by states.

5. Regulating artificial intelligence to protect human rights

These series of questions make a strong case for an IHR mandate for states to regulate AI ex-ante. We envisage a mandate in three parts: first, determining if, when, and how states will proceed to use AI systems (the 'respect' principle); second, the pre-emptive creation of rules to redress harm or an adequate reflection on whether existing rules are sufficient to address the possibly unpredictable violations derived from the use of AI technologies; and third, perhaps most complicated, to ensure that the legal regime is able to protect human rights, both in terms of the prevention of predictable harm to individuals and as a way to ensure that the intrinsic value of human rights as basic foundations of societies are guaranteed.

As noted at the beginning of the article, the IHR system is universal and binding. Human beings should have their human rights protected in any place in the world, and therefore a new AI-related IHR treaty may be necessary—the elements of such a treaty, drafted by a group of experts, are under consideration by the Council of Ministers of the Council of Europe at the time of writing[101] and excellent background work in this regard has been conducted at the Alan Turing Institute.[102] Yet from a domestic perspective there is no need to wait for a new treaty. The state is already the main guarantor of the human rights of everyone within its jurisdiction. As part of that general IHR obligation, States must consider—with a broad margin of appreciation—if and when AI-related risks are sufficiently understood to require a regulatory response.

This latter view—that states need to determine, based on their own assessments, how much regulation is needed to respect, protect, and fulfil human rights—can result in two potential approaches, both likely to be inadequate. The first, defended by many of the main actors in the AI domain, is that at this stage it is preferable to leave the field open

[100] Operational, see Tony Krone, 'International Police Operations against Online Child Pornography' (2005) 296 Trends and Issues in Crime and Criminal Justice / Australian Institute of Criminology <https://search.informit.org/doi/abs/10.3316/agispt.20053519> accessed 4 October 2021; or political, for example OSCE Parliamentary Assembly, 'Brussels Declaration - Resolution on Combating Trafficking and Exploitation of Children in Pornography' (2006) <https://www.legislationline.org/documents/id/8534> accessed 4 October 2021.

[101] Council of Europe, 'The CAHAI Held Its 6th and Final Plenary Meeting' (*Artificial Intelligence*, December 2021) <https://www.coe.int/en/web/artificial-intelligence/newsroom/-/asset_publisher/csARLoSVrbAH/content/outcome-of-cahai-s-6th-plenary-meeting> accessed 7 March 2022.

[102] David Leslie and others, 'Human Rights, Democracy, and the Rule of Law Assurance Framework for AI Systems: A Proposal' [2022] arXiv:2202.02776 [cs] <http://arxiv.org/abs/2202.02776> accessed 7 March 2022.

for self-regulation. The idea of 'disruptive technologies', first mooted in 1995,[103] has prompted ongoing debate in policy and academia over the adequacy of regulatory intervention in such technologies[104] and the risks derived from early regulatory action as an obstacle to innovation[105] or to the technology's development.[106] Regulation can in fact come at the expense of delaying or undermining the improvement of human wellbeing.[107]

Moreover, when it comes to the risks of human rights violations, an extensive body of scholarship exists, including a text co-authored by one of the authors of this paper,[108] on the potential to use self-regulatory mechanisms by business and corporations.[109] The logic behind non-compulsory reporting mechanisms and due diligence is that in due time they will give rise to new normative platforms.[110] These soft-law mechanisms and guiding principles can even act as interim measures that eventually give way to binding law in the form of treaties.[111] Yet AI presents a terrain in which the questions about risks cannot be addressed by the operators alone, and where the implications for a wider number of people appear to be unusually profound.

The second approach assumes that existing norms to protect against the worst human rights violations are sufficient. In this view, there is no current need to regulate AI; any harm could be solved within the existing regulatory regimes of responsibility or liability. In our view, although there is an element of reason in this approach, we are doubtful that the analysis it presupposes—that the system is ready to address the harms—has ever been done.[112] This is because we are yet to see sufficient examples of litigation attracting adequate forms of compensation or triggering significant changes in the ways AI systems are developed or used. Particularly in the private sector, most scandals have gone largely unaddressed from a liability point of view. In the public sector, where rules (administrative law) can be more restrictive, some litigations have been successful.[113]

[103] Joseph L Bower and Clayton M Christensen, 'Disruptive Technologies: Catching the Wave' (1995). 73 Harvard Business Review 43.

[104]

[105] Dmitrii Trubnikov, 'Analysing the Impact of Regulation on Disruptive Innovations: The Case of Wireless Technology' (2017) 17 Journal of Industry, Competition and Trade 399.

[106] Daniel Gervais, 'The Regulation of Inchoate Technologies' (2010) 47 Houston Law Review 665, 682–684.

[107] Gonenc Gurkaynak, Ilay Yilmaz and Gunes Haksever, 'Stifling Artificial Intelligence: Human Perils' (2016) 32 Computer Law & Security Review 749.

[108] Ramona Vijeyarasa and Mark Liu, 'Fast Fashion for 2030: Using the Pattern of the Sustainable Development Goals (SDGs) to Cut a More Gender-Just Fashion Sector' [2022] Business and Human Rights Journal.

[109] Geordan Graetz and Daniel M Franks, 'Incorporating Human Rights into the Corporate Domain: Due Diligence, Impact Assessment and Integrated Risk Management' (2013) 31 Impact Assessment and Project Appraisal 97.

[110] Maddalena Neglia, 'The UNGPs — Five Years on: From Consensus to Divergence in Public Regulation on Business and Human Rights' (2016) 34 Netherlands Quarterly of Human Rights 289.

[111] Claire Methven O'Brien, 'Transcending the Binary: Linking Hard and Soft Law Through a UNGPS-Based Framework Convention' (2020) 114 American Journal of International Law 186.

[112] For the limitations of this approach see Omri Rachum-Twaig, 'Whose Robot Is It Anyway?: Liability for Artificial-Intelligence-Based Robots' [2020] U. Ill. L. Rev. 1141.

[113] Jon Henley and Robert Booth, 'Welfare Surveillance System Violates Human Rights, Dutch Court Rules' *The Guardian* (5 February 2020) <https://www.theguardian.com/technology/2020/feb/05/welfare-

Another reasonable objection to this approach is one founded on moral grounds. Just as we consider criminal law a 'species of political and moral philosophy' essential to protect certain values, we need something beyond civil liability[114] to transmit the state's position regarding the limits that human rights impose on the use of AI.

A third approach, definitely more sophisticated than the other two, is to move the discussion to a different domain altogether, where private and public entities cooperate. In the words of one participants in the Glion Human Rights Dialogue cited at the beginning of this article, the 'only way to fully safeguard civil and political rights, and address threats to democracy, in the digital age, is through public-private partnership', an approach similar to the Christchurch Call to Action in the context of terrorism and violent extremism.[115] The most compelling case in this domain was put forward in the Toronto Declaration of 2018. Although its main limitation is its narrow scope—'protecting the right to equality and non-discrimination in machine learning systems'—the Declaration is a well-thought attempt to commit public and private parties to guarantee safer and fairer AI systems. However, the wording of the whole declaration can be summarized by two paragraphs that essentially exhort states to not engage in or support discriminatory practices and to protect individuals from them, including through legislation. It demands that states regulate:

> 22. States bear the primary duty to promote, protect, respect and fulfil human rights. Under international law, states must not engage in, or support discriminatory or otherwise rights-violating actions or practices when designing or implementing machine learning systems in a public context or through public-private partnerships.
> 24. States have positive obligations to protect against discrimination by private sector actors and promote equality and other rights, including through binding laws.[116]

In this respect, a growing and notable body of voices have offered the view that the *only* way to reasonably address the challenge of AI is through legislation. Quoting Alphabet CEO Sundar Pichai: 'There is no question in my mind that artificial intelligence needs to be regulated. It is too important [to] not [do it]'.[117] One cannot overstate the significance of such a statement coming from the CEO of the holding company for Google.

surveillance-system-violates-human-rights-dutch-court-rules> accessed 29 December 2021; Luke Henriques-Gomes, 'Robodebt: Court Approves $1.8bn Settlement for Victims of Government's "Shameful" Failure' *The Guardian* (11 June 2021) <https://www.theguardian.com/australia-news/2021/jun/11/robodebt-court-approves-18bn-settlement-for-victims-of-governments-shameful-failure> accessed 29 December 2021.

[114] George P Fletcher, *Rethinking Criminal Law* (Oxford University Press 2000) xix.

[115] Glion Human Rights Dialogue (n 3) 21.

[116] 'The Toronto Declaration - Protecting the Right to Equality and Non-Discrimination in Machine Learning Systems' <https://www.torontodeclaration.org/declaration-text/english/> accessed 30 September 2021.

[117] Rob Toews, 'Here Is How The United States Should Regulate Artificial Intelligence' (*Forbes*, 28 June 2020) <https://www.forbes.com/sites/robtoews/2020/06/28/here-is-how-the-united-states-should-regulate-artificial-intelligence/> accessed 10 August 2021.

Theoretically, this could be done at an international level. Erdélyi and Goldsmith, for example, advocate for the creation of an international organization with regulatory powers.[118] A more realistic approach rely on national norms, however. This does not mean there should necessarily be an AI Act 'à l'EU', designed to regulate the whole scope of AI uses through strict hard norms.[119] As succinctly explained by Chesterman in this context,[120] regulation is, in the end, a way to refer to the exercise of control (through rules, standards, or other means) by one or more public bodies, where the legitimacy lies in the connection between the means of control and a state's institutions. For example, it could be the creation of a regulatory agency that would only act when norms are needed in domain-specific areas.[121] While we are of the view that the IHR obligations of the state in relation to AI technologies can only ultimately be met through regulation,[122] in the next and final section of this article, we discuss the possibility of an interim approach as we wait for states to move in the direction of risk-based and rights-based regulation.

6. Conclusion: The role of the IHR system in enforcing a subsidiary obligation to consider the effects of AI on human rights

AI poses evident risks to the enjoyment of human rights. In this article, we have illustrated some of those risks through the lens of women's rights. In our view, these examples illustrate the inherent risks faced by other groups at risk because of particular situations of vulnerability, or for society at large. The need to address those risks require immediate state intervention. Furthermore, we believe that the absence of regulatory activity—that is, a state's failure to establish legally binding norms to protect human rights from the deployment of AI systems—is, in itself, a violation of IHR norms.

This would not be the first recognized instance under IHR law that failure to regulate would constitute a violation of the positive duties of states to protect.[123] Nonetheless, we are willing to accept that at this stage some states may not be willing or ready to regulate. In fact, as far as we are aware, the only initiative to establish a comprehensive regulation of AI is the April 2021 proposal at the European Union level, still in the early steps of its legislative process,[124] and some steps in Brazil that are more advanced in procedural terms than the EU proposal but more limited in scope.[125] We are also aware that there is

[118] Olivia J Erdélyi and Judy Goldsmith, 'Regulating Artificial Intelligence: Proposal for a Global Solution', *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (ACM 2018) <https://dl.acm.org/doi/10.1145/3278721.3278731> accessed 11 August 2021.

[119] European Commission Proposal EU AI Act (n 14).

[120] Chesterman (n 19) 3–4.

[121] Toews (n 118).

[122] Norms which should also be gender-responsive Vijeyarasa (n 95)..

[123] See the examples in Sandra Fredman, *Human Rights Transformed: Positive Rights and Positive Duties* (OUP Oxford 2008) or; Grégoire Webber, Paul Yowell and Richard Ekins, *Legislated Rights: Securing Human Rights through Legislation* (Cambridge University Press 2018).

[124] European Commission Proposal EU AI Act (n 14).

[125] Eduardo Bismarck, Bill Estabelece princípios, direitos e deveres para o uso de inteligência artificial no Brasil, e dá outras providências. 2020 [PL 21/2020].

a broader interest among some jurisdictions to proceed with certain levels or degrees of regulation, but, as noted in a 2018 scoping exercise of several major economies (selected EU countries, China, the United States, the United Kingdom, Brazil, South Africa, and others), there 'were no major or significant amendments in legislation bearing on constitutional or human rights in direct response to AI and robotics developments […] for the last five to ten years [and] [i]n some countries, even in the future this is extremely unlikely to happen'.[126]

Regardless, it is clearly necessary to start assessing which AI-related human rights challenges must be addressed by rules and enforceable requirements and which can await future human rights-based responses depending on need.[127] The above-mentioned scoping exercise noted that officials in many of the countries under study had voiced a call for action in this domain.[128] As one human rights scholar put it a few years ago, however, 'very little sustained and substantive attention has been paid to these issues by UN human rights bodies to date. In the absence of more attention at the UN level, the charge that the human rights regime is not providing much clarity and guidance to the AI debate is a valid one'.[129]

Here, we propose addressing this need through the existing mechanism of IHR systems. Following from Human Rights Council Resolution 41/11, we believe that in all the UN monitoring processes—starting with the next round of the Universal Periodic Review but extending it to all the treaty-based processes—AI must be among the list of issues that states must consider and address in their reporting. The questions noted in section four above offer a framework for doing this. An advantage of considering the question in each of the monitoring processes and committees is that asymmetric risks for different groups or types of rights will be better understood based on the experiences and knowledge of relevant experts. We hope that the treaty bodies would recommend better regulation of AI to states falling short of their obligations, but also identify particularly vulnerable groups and areas of concern as the primary subjects of such regulation.

This would serve two purposes. On the one hand, it may prompt states to acknowledge that if users and operators of AI systems move too fast and break things, they must be held responsible for the damage. Hopefully this dialogue will also prevent harms, as different stakeholders will be better placed to understand the risks of implementing AI systems in each domestic context. On the other hand, perhaps even more importantly, states will be given the opportunity to learn from each other's experiences and create

[126] Rodrigues, Siemaszko and Warso (n 5) 65.

[127] Paul Nemitz, 'Foreword · Power in Times of Artificial Intelligence' (2020) 2 Delphi - Interdisciplinary Review of Emerging Technologies 158, 8.

[128] Rodrigues, Siemaszko and Warso (n 5) 63–64.

[129] Christiaan Van Veen and Corinne Cath, 'Artificial Intelligence: What's Human Rights Got To Do With It?' (*Medium*, 18 May 2018) <https://points.datasociety.net/artificial-intelligence-whats-human-rights-got-to-do-with-it-4622ec1566d5> accessed 4 October 2021.

regulatory frameworks that suit their domestic circumstances to effectively protect the human rights of all.

Bibliography

Adams R and Loideáin NN, 'Addressing Indirect Discrimination and Gender Stereotypes in AI Virtual Personal Assistants: The Role of International Human Rights Law' (2019) 8 Cambridge International Law Journal 241

'AI for Good' (*AI for Good*) <https://aiforgood.itu.int/> accessed 2 October 2021

Aizenberg E and van den Hoven J, 'Designing for Human Rights in AI' (2020) 7 Big Data & Society 2053951720949566

Ajder H and others, 'The State of Deepfakes: Landscape, Threats, and Impact' (Deeptrace 2019) <https://regmedia.co.uk/2019/10/08/deepfake_report.pdf>

American Psychological Association, 'The Truth About Lie Detectors (Aka Polygraph Tests)' (*https://www.apa.org*, 5 August 2004) <https://www.apa.org/research/action/polygraph> accessed 30 September 2021

Australian Human Rights Commission, 'Human Rights and Technology Final Report' (2021) <https://tech.humanrights.gov.au/sites/default/files/2021-05/AHRC_RightsTech_2021_Final_Report.pdf>

Bello y Villarino J-M and Fraser H, 'Acceptable "Residual Risks" for Fundamental Rights? Understanding the Keystone of Risk-Based AI Regulations' (2021)

Boix Palop A, 'Los algoritmos son reglamentos: La necesidad de extender las garantías propias de las normas reglamentarias a los programas empleados por la administración para la adopción de decisiones' <https://repositorio.uam.es/handle/10486/692210> accessed 12 September 2021

Bower JL and Christensen CM, 'Disruptive Technologies: Catching the Wave'

Chatterjee S, N.S. S and Hussain Z, 'Evolution of Artificial Intelligence and Its Impact on Human Rights: From Sociolegal Perspective' (2021) ahead-of-print International Journal of Law and Management <https://doi.org/10.1108/IJLMA-06-2021-0156> accessed 2 October 2021

Chesterman S, *We, the Robots?: Regulating Artificial Intelligence and the Limits of the Law* (Cambridge University Press 2021)

Copps MJ, 'Disruptive Technology ... Disruptive Regulation' (2005) 2005 Michigan State Law Review 309

Council of Europe, 'The CAHAI Held Its 6th and Final Plenary Meeting' (*Artificial Intelligence*, December 2021) <https://www.coe.int/en/web/artificial-intelligence/newsroom/-/asset_publisher/csARLoSVrbAH/content/outcome-of-cahai-s-6th-plenary-meeting> accessed 7 March 2022

Davidson J, ‘Australian Law behind the Times on Deepfake Videos’ *Australian Financial Review* (8 July 2019) <https://www.afr.com/technology/australian-law-behind-the-times-on-deepfake-videos-20190703-p523pi> accessed 2 October 2021

Donahoe E and Metzger MM, ‘Artificial Intelligence and Human Rights’ (2019) 30 Journal of Democracy 115

Easton S, ‘Nadia: The Curious Case of the Digital Missing Person’ [2019] *The Mandarin* <https://www.themandarin.com.au/106473-nadia-the-curious-case-of-the-digital-missing-person/> accessed 2 October 2021

Erdélyi OJ and Goldsmith J, ‘Regulating Artificial Intelligence: Proposal for a Global Solution’, *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (ACM 2018) <https://dl.acm.org/doi/10.1145/3278721.3278731> accessed 11 August 2021

European Union Agency for Fundamental Rights, ‘#BigData: Discrimination in Data-Supported Decision Making’ (2018) <https://fra.europa.eu/en/publication/2018/bigdata-discrimination-data-supported-decision-making> accessed 30 July 2021

Expert Group on Artificial Intelligence at the OECD, ‘Scoping the OECD AI Principles: Deliberations of the Expert Group on Artificial Intelligence at the OECD’ (OECD 2019) 291 <https://read.oecd-ilibrary.org/science-and-technology/scoping-the-oecd-ai-principles_d62f618a-en> accessed 30 September 2021

Fang B, ‘Decision Support System (DSS)-Form, Development and Future’, *2009 First International Workshop on Education Technology and Computer Science* (2009)

Fletcher GP, *Rethinking Criminal Law* (Oxford University Press 2000)

Fredman S, *Human Rights Transformed: Positive Rights and Positive Duties* (OUP Oxford 2008)

Gervais D, ‘The Regulation of Inchoate Technologies’ (2010) 47 Houston Law Review 665

Glion Human Rights Dialogue, ‘Human Rights in the Digital Age: Making Digital Technology Work for Human Rights’ (Universal Rights Group 2020) <https://www.universal-rights.org/urg-policy-reports/human-rights-in-the-digital-age-making-digital-technology-work-for-human-rights/>

González Álvarez JL, López Ossorio JJ and Muñoz Rivas M, *La valoración policial del riesgo de violencia contra la mujer pareja en España – Sistema VioGén* (Ministerio del Interior Gobierno de España 2018)

González JL and Garrido MJ, ‘Satisfacción de las víctimas de violencia de género con la actuación policial en España. Validación del Sistema VioGen’ (2015) 25 Anuario de Psicología Jurídica 29

González-Prieto Á and others, ‘Machine Learning for Risk Assessment in Gender-Based Crime’ [2021] arXiv:2106.11847 [cs, stat] <http://arxiv.org/abs/2106.11847> accessed 23 September 2021

Graetz G and Franks DM, ‘Incorporating Human Rights into the Corporate Domain: Due Diligence, Impact Assessment and Integrated Risk Management’ (2013) 31 Impact Assessment and Project Appraisal 97

Gurkaynak G, Yilmaz I and Haksever G, ‘Stifling Artificial Intelligence: Human Perils’ (2016) 32 Computer Law & Security Review 749

Haenlein M and Kaplan A, ‘A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence’ (2019) 61 California management review 5

Hao K, ‘A Horrifying New AI App Swaps Women into Porn Videos with a Click’ (*MIT Technology Review*, 13 September 2021) <https://www.technologyreview.com/2021/09/13/1035449/ai-deepfake-app-face-swaps-women-into-porn/> accessed 2 October 2021

Henley J and Booth R, ‘Welfare Surveillance System Violates Human Rights, Dutch Court Rules’ *The Guardian* (5 February 2020) <https://www.theguardian.com/technology/2020/feb/05/welfare-surveillance-system-violates-human-rights-dutch-court-rules> accessed 29 December 2021

Henriques-Gomes L, ‘Robodebt: Court Approves $1.8bn Settlement for Victims of Government’s “Shameful” Failure’ *The Guardian* (11 June 2021) <https://www.theguardian.com/australia-news/2021/jun/11/robodebt-court-approves-18bn-settlement-for-victims-of-governments-shameful-failure> accessed 29 December 2021

Hulme S, Morgan A and Boxall H, ‘Domestic Violence Offenders, Prior Offending and Reoffending in Australia’ (2019) No. 580 Trends and Issues in Crime and Criminal Justice 22

Jansen P and others, ‘SIENNA D4.1: State-of-the-Art Review: Artificial Intelligence and Robotics’ (Zenodo 2019) <https://zenodo.org/record/4066571> accessed 12 August 2021

Khalid A, ‘Deepfake Videos Are a Far, Far Bigger Problem for Women’ (*Quartz*, 9 October 2019) <https://qz.com/1723476/deepfake-videos-feature-mostly-porn-according-to-new-study-from-deeptrace-labs/> accessed 3 October 2021

Kirchengast T, ‘Deepfakes and Image Manipulation: Criminalisation and Control’ (2020) 29 Information & Communications Technology Law 308

Kirchengast T and Crofts T, ‘The Legal and Policy Contexts of “Revenge Porn” Criminalisation: The Need for Multiple Approaches’ (2019) 19 Oxford University Commonwealth Law Journal 1

Knaus C, ‘NDIA Denies Cate Blanchett-Voiced “Nadia” Virtual Assistant Is in Doubt’ *The Guardian* (21 September 2017) <https://www.theguardian.com/australia-news/2017/sep/22/ndia-denies-cate-blanchett-voiced-nadia-virtual-assistant-is-in-doubt> accessed 2 October 2021

Krafft TD, Zweig KA and König PD, ‘How to Regulate Algorithmic Decision-Making: A Framework of Regulatory Requirements for Different Applications’ (2020) n/a Regulation & Governance <http://onlinelibrary.wiley.com/doi/abs/10.1111/rego.12369> accessed 30 September 2021

Krone T, 'International Police Operations against Online Child Pornography' (2005) 296 Trends and Issues in Crime and Criminal Justice / Australian Institute of Criminology <https://search.informit.org/doi/abs/10.3316/agispt.20053519> accessed 4 October 2021

Kropp PR and Hart SD, 'The Spousal Assault Risk Assessment (SARA) Guide: Reliability and Validity in Adult Male Offenders' (2000) 24 Law and Human Behavior 101

Latonero M, 'Governing Artificial Intelligence: Upholding Human Rights & Dignity' (2018) <https://apo.org.au/sites/default/files/resource-files/2018-10/apo-nid196716.pdf>

Leslie D and others, 'Human Rights, Democracy, and the Rule of Law Assurance Framework for AI Systems: A Proposal' [2022] arXiv:2202.02776 [cs] <http://arxiv.org/abs/2202.02776> accessed 7 March 2022

Liu H-Y and Zawieska K, 'A New Human Rights Regime to Address Robotics and Artificial Intelligence' [2016] JusLetter IT

Livingston S and Risse M, 'The Future Impact of Artificial Intelligence on Humans and Human Rights' (2019) 33 Ethics & International Affairs 141

Martin N, *Online Predators Spread Fake Porn of Me. Here's How I Fought Back* (2017) <https://www.ted.com/talks/noelle_martin_online_predators_spread_fake_porn_of_me_here_s_how_i_fought_back> accessed 2 October 2021

Miao Z, 'Investigation on Human Rights Ethics in Artificial Intelligence Researches with Library Literature Analysis Method' (2019) 37 The Electronic Library 914

'Mujeres - Delegación Del Gobierno Contra La Violencia de Género' (2021) <https://violenciagenero.igualdad.gob.es/violenciaEnCifras/victimasMortales/fichaMujeres/home.htm> accessed 23 September 2021

Muñoz Vicente JM and López-Ossorio JJ, 'Valoración psicológica del riesgo de violencia: alcance y limitaciones para su uso en el contexto forense' (2016) 26 Anuario de Psicologia Juridica 130

Neglia M, 'The UNGPs — Five Years on: From Consensus to Divergence in Public Regulation on Business and Human Rights' (2016) 34 Netherlands Quarterly of Human Rights 289

Nemitz P, 'Foreword · Power in Times of Artificial Intelligence' (2020) 2 Delphi - Interdisciplinary Review of Emerging Technologies 158

Ni Loideain N, Adams R and Clifford D, 'Gender as Emotive AI and the Case of "Nadia": Regulatory and Ethical Implications' (Social Science Research Network 2021) SSRN Scholarly Paper ID 3858431 <https://papers.ssrn.com/abstract=3858431> accessed 6 September 2021

O'Brien CM, 'Transcending the Binary: Linking Hard and Soft Law Through a UNGPS-Based Framework Convention' (2020) 114 American Journal of International Law 186

OECD, *Artificial Intelligence in Society* (OECD 2019) <https://www.oecd-ilibrary.org/science-and-technology/artificial-intelligence-in-society_eedfee77-en> accessed 7 June 2021

Office of the High Commissioner for Human Rights, ‘Gender Stereotyping as a Human Rights Violation’ (2013)

OHCHR, ‘A Human-Rights-Based Approach to Data’ <https://www.ohchr.org/Documents/Issues/HRIndicators/GuidanceNoteonApproachtoData.pdf> accessed 19 July 2018

——, ‘Artificial Intelligence Ensuring Human Rights at the Heart of the Sustainable Development Goals’ (10 March 2021) <https://www.ohchr.org/EN/NewsEvents/Pages/ArtificialIntelligence-SDGs.aspx> accessed 12 July 2021

OSCE Parliamentary Assembly, ‘Brussels Declaration - Resolution on Combating Trafficking and Exploitation of Children in Pornography’ <https://www.legislationline.org/documents/id/8534> accessed 4 October 2021

Pasquale F, *New Laws of Robotics. Defending Human Expertise in the Age of AI* (Belknap Press 2020)

Paul K, ‘California Makes “Deepfake” Videos Illegal, but Law May Be Hard to Enforce’ *The Guardian* (7 October 2019) <https://www.theguardian.com/us-news/2019/oct/07/california-makes-deepfake-videos-illegal-but-law-may-be-hard-to-enforce> accessed 3 October 2021

Penney J and others, ‘Advancing Human Rights-by-Design in the Dual-Use Technology Industry’ (2018) 20 Columbia Journal of International Affairs <https://digitalcommons.schulichlaw.dal.ca/scholarly_works/250>

Pinedo M, ‘Matemáticas e inteligencia artificial contra el maltrato machista’ *EL PAÍS* (2 September 2021) <https://elpais.com/sociedad/2021-09-02/matematicas-e-inteligencia-artificial-contra-el-maltrato-machista.html> accessed 23 September 2021

Probyn E by political editor A, ‘Government’s Blanchett-Voiced AI Venture for NDIS Stalls’ *ABC News* (21 September 2017) <https://www.abc.net.au/news/2017-09-21/government-stalls-ndis-virtual-assistant-voiced-by-cate-blanchet/8968074> accessed 25 September 2021

Rachum-Twaig O, ‘Whose Robot Is It Anyway?: Liability for Artificial-Intelligence-Based Robots’ [2020] U. Ill. L. Rev. 1141

Randhavane T and others, ‘Identifying Emotions from Walking Using Affective and Deep Features’ (2020) <http://arxiv.org/abs/1906.11884> accessed 1 October 2021

Raso FA and others, ‘Artificial Intelligence & Human Rights: Opportunities & Risks’ (Social Science Research Network 2018) SSRN Scholarly Paper ID 3259344 <https://papers.ssrn.com/abstract=3259344> accessed 23 September 2021

Ray A, ‘Disinformation, Deepfakes and Democracies: The Need for Legislative Reform’ (2021) 44 UNSW Law Journal 983

Renda A, ‘Europe: Toward a Policy Framework for Trustworthy AI’, *The Oxford Handbook of Ethics of AI* (2020)

Risse M, ‘Human Rights and Artificial Intelligence: An Urgently Needed Agenda’ (2019) 41 Human Rights Quarterly 1

Rodrigues R, Siemaszko K and Warso Z, ‘SIENNA D4.2: Analysis of the Legal and Human Rights Requirements for AI and Robotics in and Outside the EU’ (Zenodo 2019) <https://zenodo.org/record/4066812> accessed 12 August 2021

Schwarz EC, ‘Human vs. Machine: A Framework of Responsibilities and Duties of Transnational Corporations for Respecting Human Rights in the Use of Artificial Intelligence Notes’ (2019) 58 Columbia Journal of Transnational Law 232

Singer P, ‘Morality, Reason, and the Rights of Animals’, *Primates and Philosophers: How Morality Evolved* (Princeton University Press 2009) <http://www.degruyter.com/document/doi/10.1515/9781400830336-010/html> accessed 1 October 2021

Smith R, ‘Nadia Falters: Teetering Technology in the Service of Access to Justice’ (*Law, Technology and Access to Justice*, 6 November 2017) <https://law-tech-a2j.org/advice/nadia-falters-teetering-technology-in-the-service-of-access-to-justice/> accessed 2 October 2021

State Council of China, ‘China’s New Generation of Artificial Intelligence Development Plan (Non-Official Translation)’ <https://flia.org/notice-state-council-issuing-new-generation-artificial-intelligence-development-plan/> accessed 7 October 2021

Sunstein CR, ‘Maximin’ (2020) 37 Yale Journal on Regulation 940

‘The Toronto Declaration - Protecting the Right to Equality and Non-Discrimination in Machine Learning Systems’ <https://www.torontodeclaration.org/declaration-text/english/> accessed 30 September 2021

Toews R, ‘Here Is How The United States Should Regulate Artificial Intelligence’ (*Forbes*, 28 June 2020) <https://www.forbes.com/sites/robtoews/2020/06/28/here-is-how-the-united-states-should-regulate-artificial-intelligence/> accessed 10 August 2021

Tomuschat C, *Human Rights: Between Idealism and Realism* (OUP Oxford 2014)

Trubnikov D, ‘Analysing the Impact of Regulation on Disruptive Innovations: The Case of Wireless Technology’ (2017) 17 Journal of Industry, Competition and Trade 399

UNESCO, ‘Recommendation on the Ethics of Artificial Intelligence’ <https://unesdoc.unesco.org/ark:/48223/pf0000380455> accessed 7 March 2022

Van Hoey J and others, ‘Profile Changes in Male Partner Abuser After an Intervention Program in Gender-Based Violence’ (2021) 65 International Journal of Offender Therapy and Comparative Criminology 1411

Van Veen C and Cath C, ‘Artificial Intelligence: What’s Human Rights Got To Do With It?’ (*Medium*, 18 May 2018) <https://points.datasociety.net/artificial-intelligence-whats-human-rights-got-to-do-with-it-4622ec1566d5> accessed 4 October 2021

Vijeyarasa R, ‘CEDAW’s General Recommendation No. 35: A Quarter of a Century of Evolutionary Approaches to Violence against Women’ (2020) 19 Journal of Human Rights 153

——, ‘In Pursuit of Gender-Responsive Legislation: Transforming Women’s Lives through the Law’ in Ramona Vijeyarasa (ed), *International Women’s Rights Law and Gender Equality: Making the law work for women* (Routledge, Taylor and Francis 2021)

Vijeyarasa R and Liu M, ‘Fast Fashion for 2030: Using the Pattern of the Sustainable Development Goals (SDGs) to Cut a More Gender-Just Fashion Sector’ [2022] Business and Human Rights Journal

Wagner TL and Blewer A, ‘“The Word Real Is No Longer Real”: Deepfakes, Gender, and the Challenges of AI-Altered Video’ (2019) 3 Open Information Science 32

Waldman AE, ‘Power, Process, and Automated Decision-Making Symposium: Rise of the Machines: Artificial Intelligence, Robotics, and the Reprogramming of Law’ (2019) 88 Fordham Law Review 613

Webber G, Yowell P and Ekins R, *Legislated Rights: Securing Human Rights through Legislation* (Cambridge University Press 2018)

Williams PJ, *The Alchemy of Race and Rights* (Harvard University Press 1991)

Wise SM, ‘A Great Shout: Legal Rights for Great Apes’, *Animal Rights* (Routledge 2008)

Wogu IAP and others, ‘Human Rights’ Issues and Media/Communication Theories in the Wake of Artificial Intelligence Technologies: The Fate of Electorates in Twenty-First-Century American Politics’ in Thangaprakash Sengodan, M Murugappan and Sanjay Misra (eds), *Advances in Electrical and Computer Technologies* (Springer 2020)

Xenidis R and Senden L, ‘EU Non-Discrimination Law in the Era of Artificial Intelligence: Mapping the Challenges of Algorithmic Discrimination’ (Social Science Research Network 2019) SSRN Scholarly Paper ID 3529524 <https://papers.ssrn.com/abstract=3529524> accessed 25 September 2021

Yeung K, Howes A and Pogrebna G, ‘AI Governance by Human Rights–Centered Design, Deliberation, and Oversight’, *The Oxford Handbook of Ethics of AI* (Oxford University Press 2020)

Zalnieriute M, Moses LB and Williams G, ‘The Rule of Law and Automation of Government Decision-Making’ (2019) 82 The Modern Law Review 425

Završnik A, ‘Criminal Justice, Artificial Intelligence Systems, and Human Rights’ (2020) 20 ERA Forum 567

Zeng J, ‘Artificial Intelligence and China’s Authoritarian Governance’ (2020) 96 International Affairs 1441

Bismarck E, Bill Estabelece princípios, direitos e deveres para o uso de inteligência artificial no Brasil, e dá outras providências. 2020 [PL 21/2020]

CEDAW Committee, General Recommendation No. 19, Violence against women (Eleventh Session, 1992) 1992

——, General recommendation No. 35 on gender-based violence against women, updating general recommendation No. 19 2017 [UN. Doc. No. CEDAW /C/GC/35]

European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts 2021

Convention on the Rights of Persons with DIsabilities 2006

Convention relating to the Status of Refugees 1951

Crimes Act 1900 No 40 - NSW

Ley Orgánica 1/2004, de 28 de diciembre, de Medidas de Protección Integral contra la Violencia de Género. 2004 (BOE-A-2004-21760)