A NEW HUMAN RIGHTS REGIME TO ADDRESS ROBOTICS AND ARTIFICIAL INTELLIGENCE

Hin-Yan Liu / Karolina Zawieska

Associate Professor, University of Copenhagen, Centre for International Law, Conflict and Crisis, Faculty of Law
Karen Blixens Plads 16, 2300 Copenhagen, DK
hin-yan.liu@jur.ku.dk

Researcher, Industrial Research Institute for Automation and Measurements PIAP, Fundamental Research Team
Al. Jerozolimskie 202, 02-486 Warsaw, PL
kzawieska@piap.pl

Keywords: ***New Human Rights Regime, Defining Human Values, «Man-under-the-loop», AI power***

Abstract: *The relationship between human beings and AI appears to be on the cusp of foundational change: until recently the human ability to control technology was unquestioned, but now the trend suggests a declining power differential and the possibility of an inverse power relationship soon. AI is poised to exert increasing influence over human opportunities and activities, such that human beings are increasingly under 'the loop'. This paper explores the impact that the inversion of power between human beings and their technologies has on the protection of fundamental human rights.*

1. Introduction[1]

The capabilities of contemporary weak artificial intelligence (AI) are thought to have superseded those of human experts in a range of narrow activities, and have been vaunted to outpace human performance in expanding fields of endeavour in the near future. Thus, even under existing circumstances, the relationship between human beings and their technologies appears to be on the cusp of a foundational change: until recently the human ability to control technology was unquestioned, but now the trend leans towards a declining power differential thereby setting the trajectory towards the inversion of this power relationship. While these effects are currently only discernible within limited domains of human activity, it is worth noting both that the impact of weak AI is deepening and that its scope is broadening. In other words, as AI is poised to exert increasing influence and power over human opportunities and activities in ever broader spheres: in the parlance of *the loop*, human beings may no longer be *in*, *on*, or *out of*, the loop but are instead increasingly *under* that loop. Our overarching claim in this paper is that regulatory strategies need to be devised now that are capable of taking into account the prospect of a power reversal between human beings and AI, and our broad aim is to ensure the continuation of protections for human persons and society in the face of technological, rather than political, economic or military power.

This paper explores the impact that the inversion of power between human beings and their technologies has on the protection of fundamental human rights. While it is significant that the use of weak AI is already the source of challenges to existing human rights protections today, these challenges have been articulated as issues that are contained within certain spheres of activities or in relation to enumerated legal protections. We argue that reliance upon such an orthodox approach to human rights and AI may miss the mark. This creates a false sense of security in the relationship between human beings and the effects of AI by failing to recognise the increasing power these entities exercise over human persons, leaving technological determinism unquestioned.

[1] This work is a concise version of a paper currently under review with Ethics and Information Technology.

To advance this argument, we first dispatch with some preliminary objections before proceeding to highlight three structural obstacles latent within the existing human rights regime that bar the way to developing a human rights regime against AI technologies. These are complemented by three further problems inherent within the nature of the existing human rights regime that further limit its effectiveness in relation to AI. Together these suggest significant shortcomings that need to be redressed before widespread societal adoption of AI. Taking this structural perspective of the relationship between human rights and AI counterbalances what isolated perspectives obfuscate. We question whether such a fragmented perspective, where each challenge in a given domain is confronted independently, can capture the true nature of what might be termed a broader technological threat to human rights. Viewed in gestalt terms, where the whole is different than the sum of the parts, the entire challenge to human rights protection posed by AI is fundamentally different in nature to those human rights erosions that are visible through the orthodox legal lens. This suggests that an alternative approach, centred upon the power reversal between human beings and AI and focussing upon the relational dimension, needs to be developed in order to maintain sufficient human rights protection moving forward. By panning out the perspective, the opportunities to enhance human rights protections both in relation to and through technology may also be considered.

2. Initial Objections

The proposal to re-orientate the human rights regime against technological incursions will confront an initial objection that the locus of power remains inherent within the State and its institutions such that increased capabilities leveraged through weak AI merely enhance intelligence and collectivise capacity of human-computer teams. Phrased differently, the objection is that erosions of human rights protections are at the second order usage of AI, rather than first order challenges posed directly by the technologies themselves. The objection is therefore that there is no need for human rights to be orientated against technology because the orthodox rights mechanisms can be deployed against the human organisations that remain at the core of any human rights infringement.

Our rejoinder is two-fold: that the impetus towards a human rights regime against technology in no way implies a relegation of the existing system of human rights protections against the State; and that human rights laws are but one of an array of 'Swiss cheese' obstacles against the occurrence of violations [Reason 2000]. Our articulation of a human rights regime against technology is envisaged to reinforce the spirit of the movement and to ensure continued protections against new types of powerful threats. Thus, even if weak AI merely enhances the hand of the State in relation to the individual, the homeostatic equilibrium between State power and human rights would be upset because the contemporary legal constellation is predicated upon the continued existence and efficacy of complementary restrictions to the exercise of power which is eroded by the technology.

3. Three Structural Obstacles

Even if the challenge posed by weak AI to the protection of human rights is conceded, additional obstacles need to be cleared before the path is clear to develop human rights protections against technology can be explored. The first is the tendency to compartmentalise concerns according to the sphere of activity or the nature of the impugned right [Kennedy 2005]. This tendency towards isolated considerations leads to a fragmented understanding of the true nature of the problem as a whole. The possibility that a larger structural shift is taking place is hidden by the fact that an incomplete portrait has been painted, and this truncated understanding militates against progressive developments of human rights protections.

The second is that contemporary human rights methodologies are extremely effective at illuminating certain enumerated types of harms caused by the State and its agents to identified individual victims within jurisdictional boundaries. The efficiency of this mechanism, however, risks leaving unrecognised harms that fall outside of this formula [Veitch 2007]. This essentially amounts to distinguishing between legitimate and ille-

gitimate forms of harm, and sterilises technologically induced harms from the stigma of human rights abuse. Not only does this render ineffective avenues of remedy and redress against power wielded through technological means, but excludes an increasingly powerful agency from the purview of review and responsibility. As discussed below, this is an inherent issue subsisting within human rights law.

Finally, there is the monopolising tendency of human rights law that crowds out other perspectives on the pertinent issues [Kennedy 2005]. The consequence of this hegemonising pressure is really that any claim to defend human values must be couched within its logic and language to be successful. This obstacle imposes significant constraints on the possibility of deviance away from the dominant human rights model, despite the cardinal features of this model being inherent to the problem in the first place.

4. The Interface between Contemporary Human Rights and Emerging AI

The problems inherent within the contemporary configuration of human rights law that curtail its effectiveness in relation to what might be termed technological power (as opposed to State power) are threefold. First, there is the content of existing human rights law: the substantive rights have largely evolved in relation and against State power borne out in the experiential theory of human rights [Dershowitz 2009]. This suggests that technological wrongs are required before appropriate technologically-oriented rights can emerge as a reaction. More problematically, however, is that the substance of existing rights are not aligned with the challenges posed by AI: the freedoms of speech and assembly, for example, are tailored to resist against State repression but may not fully overlap with the concerns raised by our emerging technologies. Second, human rights law, ossified within the State-orientated approach, renders it oblivious to all other power dynamics that potentially impact human beings and challenge the very concept of the human individual. Not only does this overlook first order challenges raised by AI directly, but also gives rise to complex second order problems where corporations, for example, deploy AI that would erect two interlocking jurisdictional barriers to traditional human rights claims. Finally, the legal focus upon isolated direct causal relationships raises the third problem because dispersed or distributed origins of harms and indirect or tangential effects cannot be recognised within this framework [Isaacs & Vernon 2011; Nollkaemper & van der Wilt 2009]. As the impact of AI is likely to arise cumulatively, its disparate effects will only be discernible through a broadened perspective, such that technological harms will fail to be recognised.

5. The Need for a Human Rights Regime Oriented Against AI Power

Having identified the perils of a power inversion between human beings and developing AI, we argue that an appropriately aligned regime needs to be developed to confront technological power directly to ensure the continuity of human rights protection. Building this regime upon the human rights discourse allows for both the refinement and reassertion of core human values in the face of technological challenges. As AI forces deep and critical re-evaluations of what it means to be human, steps that are capable of strengthening the individual in the face of technological incursions are necessary to provide adequate protection for human beings.

In eroding the traditional human rights linkage between the State and the individual, and refocusing instead on the raison d'etre of pushing back against power, forging a relational connection between human rights and technological power, as manifested through AI, opens space for rights-based mechanisms to challenge other powerful entities such as private corporations that have largely deflected such obligations.

6. Advantages of a Human Rights Regime Oriented Towards AI Challenges

The effort of devising a convergent human rights regime that is directed specifically against technological power, manifested in this case by robotics and AI, which can be asserted where situations fall into the responsibility gap [Matthias 2004]. Building a complementary human rights regime holds forth the benefit of balancing responsibilities and calibrating capacities: unilateral thrusts of human responsibility behind robotic

systems risk scapegoating human beings [Liu 2016], or exposes human beings as moral crumple zones where the human in a robotic system bears the responsibility for the failure of a broader system [Elish 2016].

The logic of a human rights regime against robotics levies four other advantages that supplement the drive towards responsible robotics. First, the rights approach centres upon the defence of the vulnerable party in a power relationship that is unhinged to the specific nature of the threat. In other words, the rights orientation is consequentialist in that its aegis is effective against infringements as they arise while the responsibility approach is procedural such that issues are neutralised insofar as the necessary considerations have been addressed. The approach of responsibility is simultaneously preventative and retributive: responsibilities operate before and after an event, and in any case its functioning does not catch every instance of wrong or harm. Thus, the defensiveness of the rights approach complements the restraint orientation of responsibility practices where threats nevertheless emerge.

Second, rights mechanisms can be deployed as procedures for monitoring compliance to responsibility obligations. Asserting a rights infringement can be an extremely effective avenue for uncovering the failure of responsibility practices because it broadens the range of reviewers to the entire class of potential victims. While imperfect, granting standing to challenge robotic harms may also help to refine the precise nature and contours of the responsibilities that are borne by those involved in the design, development and deployment of robotics. In this sense, the iterative processes inherent within litigation will eventually balance the practical interests in the use of robotics against the harms that they pose to individuals and society at large.

Third, despite the inherent and inalienable nature of human rights that has been propounded in international law (Universal Declaration of Human Rights), the contemporary human rights regime is essentially relational. In other words, the theory of human rights as integral to the individual is incongruous with practical human rights protections which allow these rights to be asserted only against narrowly construed sets of actors, namely the state and its agents. In this context, the process of devising a human rights regime against robotics will be more faithful to its intrinsic nature, orientated to protecting the human being against certain types of infringements irrespective of the nature or character of the source of the violations.

Fourth, the emphasis upon a human rights regime recognises the potential for an inversion of the power relationship between human beings and robotics. While we may not yet be at the stage where our robotics is beyond our control or influence, the power disparity is arguably lessening in ever widening areas of human activity. In the parlance of *the loop*, human beings may no longer be *in*, *on*, or *out of*, the loop but are instead increasingly *under* that loop. Deploying the logic of human rights recognises the prospect for such a reversal, an insofar as rights and responsibilities are a negotiated equilibrium, maximal protection of human interests will be retained the earlier this negotiation is initiated.

7. Responsibility, Control and Relationality

Given the increasing risk of leaving the human being under the loop when developing robotics and AI, the key concern is that of control. One of the main reasons for people to feel threated when confronted with robotics and AI creations is that they have only limited possibilities to control such technologies. From the perspective of individual users, the lack of control is due to various factors: limited understanding of how a given system is made and how it works; the design of the systems that often limits the possibility for external intervention; as well as an increasing degree of autonomy different systems and their functions are endowed with. The role of the individual is to act mainly as the consumer who can *use* different products and services. This includes adapting a system to his or her preferences, to a varying degree, which may give an illusion of but not the actual control over a system. When analysed from the perspective of the system designers, a reason for concern is that, as the systems become increasingly autonomous and intelligent as well as capable of learning, no one controls their conduct and the corresponding consequences. This is part of a broader socio-cultural context where neither experts nor the institutions are in a position to define and control different

risks that have emerged in the contemporary societies. And yet, 'society more than ever relies and insists on security and control' [Beck 2006, 335]. Therefore, we face a significant degree of complexity as well as contradictory trends: on the one hand, we assign an increasing degree of autonomy to robotic systems and AI as they seem to be more efficient than humans under certain aspects, and more reliable, for example in warfare [Arkin 2009]; on the other hand, the lack of or only limited control is exactly the reason for concern. Also, the control issue is directly related to the question of responsibility, including in the context of robotics: 'a person can be held responsible for something only if that person has control over it' [Marino & Tamburrini 2006, 49]. Part of such thinking is an assumption that a person can be held accountable for a given artefact to the extent he or she can foresee related risks and consequences. Thus, predictability is of crucial importance for the legal approaches to liability. At the same time, foreseeing outcomes and risks is increasingly difficult for autonomous and learning robots and AI [Asaro 2015], and hence, the lack of control and 'a responsibility gap' [Marino & Tamburrini 2006].

As discussed above, there are different types of responsibility. The underlying assumption in this work is that responsibility is *relational* in nature. While relational responsibility has been sometimes addressed in terms of a relation between people and events or consequences [Dworkin 2011, 102], other approaches, such as symbolic interactionism, emphasise the constructive nature of responsibility, where 'the assessment of responsibility always includes a process of negotiation' [Scheff 2009, 116]. In other words, the assignment of responsibility is a matter of negotiation among interactants [Walton 1985] rather than a matter of mere application of rules and norms [Becker & McCall 2009, 133]. This is related to the fact that responsibility implies both being responsible 'for something' and 'to someone'. The latter requires not only acknowledging an entity an act of responsibility is directed to but also addressing such an entity as an actor actively engaged in the process of responsibility assignment. In other words, responsibility implies *responding* to others rather than merely reacting to a given person, event or a consequence. From this perspective, responsibility requires a degree of interaction and reciprocity, where all actors have a sufficient degree of autonomy and capabilities to enter interaction and the related process of negation of meanings (mutual engagement is also relevant for rights, where '[r]ights can be seen be viewed as instituting and fostering relationships of reciprocity and interdependence' [Marshall 2014, 79]). In line with such thinking, responsibility may be *assumed* (accepted) rather than only assigned (imposed) to a given person, just as rights need to be respected rather than only prescribed. This leads us to another key component of the responsibility concept, namely the conceptualisation of responsibility as an *ability*. While responsibility may also be defined as a virtue [Dworkin 2011, 102], we argue here that it is more of a *process* rather than an attribute. This is how one may learn to be responsible, rather than is responsible (a difference clearly shown between children and adults), just as he or she may learn to respond and interact socially with others, as well as negotiate socially constructed meanings.

Therefore, responsibility goes far beyond the question of control, where controlling robots and foreseeing their conduct is in any case an increasingly challenging task. We argue here we should move from the control-related concept of responsibility towards responsibility viewed as a relational and dynamic process. One the one hand, such understanding of responsibility may prove difficult to translate into the engineering and computer scientists terms; on the other hand, it allows accommodating new forms of responsibility and disruptions from new technologies and allows developing alternative approaches centred upon the relational dimension. This shift from static to dynamic and relational conceptions of responsibility and control may serve as indicators as to the types of reorientation and reframing that would be necessary for a new rights-based regime to confront the challenges posed by robotics and AI. Given the ossified nature of traditional juridical concepts epitomised by rights and responsibilities, however, the challenge will be to overturn this inertia to ensure the continuing relevance of legal mechanisms in protecting human beings into the future.

8. Conclusions

Drawing together the increasing inadequacies of the contemporary human rights regime, the advantages of retaining and reorienting rights-based mechanisms, and the growing gap in control and responsibility introduced by robotics and AI, the need for developing a new human rights regime to ensure continued and sustained protections to the human being cannot be made more clearly. Not only are the orthodox approaches becoming increasingly tangential, perhaps towards the point of distraction, but as robotics and AI intercede between clear cause and effect pathways the foreseeability of harm becomes increasingly opaque. The uncoupling of cause and consequence through robotics and AI creates a conundrum for the responsibility mechanisms because these are modelled upon different presumptions that these technologies erode. Without a strong and dedicated rights-based mechanism to substitute for this weakness, however, the continued protection of human rights will be structurally, albeit subtly, diluted.

A different way of approaching this issue is by appealing to James Reason's 'Swiss cheese' model [Reason 2000]: the need to establish regulatory redundancy is clear if catastrophic regulatory failure is to be avoided. Our proposal to complement the responsible robotics project aims to duplicate the critical functions of the regulatory system to increase reliability, embryonic and imperfect as it is. We argue here that human rights should be developed in a way to protect humans against the outcomes of robotics and AI through strengthening the very notion of the human being as well as human value. How to achieve such a goal, remains an open question and the aim of this paper is to try to begin such discussions.

9. References

Arkin, Ronald R., Governing Lethal Behavior in Autonomous Robots, CRC Press, 2009.

Asaro, Peter M., The Liability Problem for Autonomous Artificial Agents, AAAI, 2016.

Beck, Ulrich, Living in the world risk society, Hobhouse Memorial Public Lecture given on Wednesday 15 February 2006 at the London School of Economics, Economy and Society, 2006, 35(3), pp. 329–345.

Becker, Howard S./McCall, Michal M. (eds.), Symbolic Interaction and Cultural Studies, University of Chicago Press, 2009.

Dershowitz, Alan, Rights from Wrongs: A Secular Theory of the Origins of Rights, Basic Books, 2009.

Dworkin, Ronald, Justice for Hedgehogs, Harvard University Press, 2011.

Elish, Madeleine Clare, Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction, WeRobot 2016 Working Paper, University of Miami, 2016.

Isaacs, Tracy/Vernon, Richard, Accountability for Collective Wrongdoing, Cambridge University Press, 2011.

Kennedy, David, The Dark Sides of Virtue: Reassessing International Humanitarianism, Princeton University Press, 2005.

Liu, Hin-Yan, Refining Responsibility: Differentiating Two Types of Responsibility Issues Raised by Autonomous Weapons Systems. In: Bhuta, Nehal/Beck, Susanne/Geiss, Robin/Liu, Hin-Yan/Kress, Clauss (eds.), Autonomous Weapons Systems: Law, Ethics, Policy, Cambridge University Press, 2006, pp. 325–344.

Marino, Dante/Tamburrini, Guglielmo, Learning robots and human responsibility, International Review of Information Ethics, 2006, 6(12), pp. 46–51.

Marshall, Jill, Human Rights Law and Personal Identity, Routledge, 2014.

Matthias, Andreas, The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata, Ethics and Information Technology, 2004, 6(3), pp. 175–183.

Nollkaemper, André/van der Wilt, Harmen (eds.), System Criminality in International Law, Cambridge University Press, 2009.

Reason, James, Human error: models and management, British Medical Journal, 2000, 320(7237), pp. 768–770.

Scheff, Thomas J., Being Mentally Ill: A Sociological Theory, Third Edition, Transaction Publishers, 2009.

Veitch, Scott, Law and Irresponsibility: On the Legitimation of Human Suffering, Routledge Cavendish, Oxford, 2007.

Walton, Marsha D., Negotiation of responsibility: Judgments of blameworthiness in a natural setting, Developmental Psychology, 1985, 21(4), pp. 725–736.