AI Challenges and the Inadequacy of Human Rights Protections

Liu, Hin-Yan

Published in:
Criminal Justice Ethics

DOI:
10.1080/0731129X.2021.1903709

Publication date:
2021

Document version
Publisher's PDF, also known as Version of record

Citation for published version (APA):
AI Challenges and the Inadequacy of Human Rights Protections

Hin-Yan Liu

To cite this article: Hin-Yan Liu (2021): AI Challenges and the Inadequacy of Human Rights Protections, Criminal Justice Ethics, DOI: 10.1080/0731129X.2021.1903709

To link to this article: https://doi.org/10.1080/0731129X.2021.1903709

Published online: 23 Mar 2021.
AI Challenges and the Inadequacy of Human Rights Protections

HIN-YAN LIU

My aim in this article is to set out some counter-intuitive claims about the challenges posed by artificial intelligence (AI) applications to the protection and enjoyment of human rights and to be your guide through my unorthodox ideas. While there are familiar human rights issues raised by AI and its applications, these are perhaps the easiest of the challenges because they are already recognized by the human rights regime as problems. Instead, the more pernicious challenges are those that have yet to be identified or articulated, because they arise from new affordances rather than directly through AI modeled as a technology. I suggest that we need to actively explore the potential problem space on this basis. I suggest that we need to adopt models and metaphors that systematically exclude the possibility of applying the human rights regime to AI applications. This orientation will present us with the difficult, intractable problems that most urgently require responses. There are convincing ways of understanding AI that lock out the very possibility for human rights responses and this should be grounds for serious concern. I suggest that responses need to exploit both sets of insights I present in this paper: first that proactive and systematic searches of the potential problem space need to be continuously conducted to find the problems that require responses; and second that the monopoly that the human rights regime holds with regards to addressing harm and suffering needs to be broken so that we can deploy a greater range of barriers against failures to recognize and remedy AI-induced wrongs.

Keywords: artificial intelligence, human rights, problem-finding, decision-making, designed environments, tailored environments

Introduction

By the second decade of the new millennium, we are all familiar with the purported challenges AI applications pose to the protection and enjoyment of human rights. The examples are legion, and, as they vie for media headlines, each one becomes rapidly replaced by the newest set of problems. The tension between the persistent challenges posed by AI applications to human rights, and
the perpetually shifting problems presented in the real world indicates an underlying relationship between root-causes and manifest symptoms. Assuming that I have correctly diagnosed this problem, I will endeavor to confront the root-causes and not to be distracted by the symptoms that draw our attention. For this reason, this paper will remain at an abstract and theoretical level and will be largely devoid of examples. My aim is thus to set out a general argument as to why the contemporary human rights regime is unable or insufficient to contest or contain the threats, harms, and challenges posed by AI applications.2 My hope is that, having presented the root-cause conditions for the inadequacies of human rights protections in this area, my argument can be readily applied to particular symptomatic manifestations that vex researchers and policy makers.

My focus on the capacity and capability of the human rights regime to resist and repel the challenges posed by AI applications is based on the monopoly of the human rights discourse and doctrine over the recognition of recompense for harm and suffering.3 In other words, my attempt at dissecting the failures of human rights in this paper is not intended as an assault on its general importance or efficacy. Rather it is the exact opposite impulse that fuels my concern: that the contemporary human rights regime may be brittle in the face of AI challenges. This brittleness might not necessarily be problematic on its own: often the strongest objects are the most brittle.4 Yet, coupled with its crowding out of competing frameworks for understanding human suffering, the failure of human rights protections may become catastrophic insofar as new forms and manifestations of harm that are brought about or revealed by AI applications may be overlooked or are not recognized as forms of human suffering.

The failure of recognition is critical, for the simple reason that if a form of harm or suffering cannot be captured within the human rights discourse, it will necessarily fail to initiate its mechanisms and thereby be excluded from the possibility of being a human rights violation. I will unpack this idea in detail below, but for now it is worth recognizing that an aspect of the brittleness of the human rights regime lies within its exclusionary nature: that not all wrongs, harms, or forms of suffering reach the mantle of human rights protection. Thus, the human rights mechanisms function simultaneously as both the litmus test of severe human wrongs as well as the yardstick by which to measure and censure these wrongs. Once we start bringing these ideas together, there are clear difficulties in actually identifying the human rights challenges presented by AI applications! This is a bizarre and unsettling conclusion that I will seek to unravel in this paper.

These are big claims to make, and there are important ramifications of my argument. I will first set out the tautology at play whereby the human rights regime is at once the litmus test, the benchmark, and the oversight and enforcement mechanism. This leads to the troubling conclusion that human rights violations are “nothing more or less” than a violation of human rights. Having set up the controversial claim that grounds my argument that the human rights regime will be inadequate in the face of AI challenges, I will take us
through more familiar territory: first by mapping the contours of the contemporary human rights regime, and then taking us to explore the initial shortcomings of this regime when confronted with the challenges posed by AI applications.

Beyond this familiar territory, however, more insidious dangers lurk. To identify and grapple with these perils posed by AI applications requires us to adopt new perspectives and approaches, and to get creative about the metaphors and models that we use to understand what “AI is, how AI operates, how we treat or relate to AI technology, how we use AI in society, or what (unintended) impacts AI will have on our society down the road.” It is only when we adopt such problem-finding perspectives that we can begin to see the myriad of subtle challenges that AI applications introduce or reveal. As the contemporary human rights regime is squarely a problem-solving framework, insofar as it seeks to recognize and remedy historical and orthodox forms of human suffering, it must be complemented by an intentional exploration of the potential problem-space.

Thus, when it comes to AI challenges, our collective reliance upon human rights protections may be unwarranted at best, and counter-productive at worst. In particular, I would like to suggest that our reflexive reliance upon the human rights regime overlooks the challenges brought about by the introduction of AI applications to the regime’s foundational presumptions and over-emphasizes its procedures and protective capacities. It may be prudent to reinforce the human rights regime with other protective principles, procedures, and processes in order to minimize the possibility that harm and suffering go unrecognized and unresolved.

The “Purported” Challenges Posed by AI to the Human Rights Regime

At the outset, I would like to emphasize the term “purported” in the section heading: the purported challenges AI applications pose to the protection and enjoyment of human rights. This is a useful way to unpack some of the issues set out in this paper. First, there is an obvious tautological quality involved in deploying human rights as the benchmark by which AI promises and perils are measured. If the aim is merely to arrive at human rights compliance for AI applications, then clearly this problem vanishes, because the adequacy and appropriateness of the yardstick remains unquestioned.

Yet, as my claim is that the contemporary human rights regime is inadequate to protect against the challenges raised or revealed by AI applications, I call into question the very metric of measurement itself. The implication is that other indicators and metrics will be required to fully gauge the nature and scope of the challenges posed by AI applications. In slogan form: human rights compliance may be necessary, but mere compliance with the requirements imposed by the contemporary human rights regime will be insufficient.

Second, reliance upon the human rights framework as the yardstick...
against which AI applications are measured implies that the human rights regime is a complete enterprise. The implicit conclusion is that the human rights regime is “complete” in the sense that it comprehensively covers the entire territory of human wrongs. This is a subtle point, but at minimum it suggests that we know what the problems are, what the solution looks like, and the steps that we need to take to get from here to there. A different way of approaching this issue is to think of the human rights framework as static, established, and entrenched. After all, the defining quality of a yardstick is that it is a reliable and consistent measure of physical length.

An important corollary of the implied completeness of the contemporary human rights regime is that it directly conflicts with the “living instrument” doctrine. This is an “evolutive” doctrine within the jurisprudence of the European Court of Human Rights (ECtHR) through which the ECtHR interprets the European Convention of Human Rights in light of present day conditions. The evolution of the human rights protection standard over time contradicts any implicit presumption that the human rights framework is a complete enterprise. While I do concede that the “living instrument” doctrine permits latitude only in the interpretation of the Convention, with the implication that the substantive protections remain fixed, the point is that the efficacy of human rights protections need to take account of changes in the sociotechnical landscape.

Third, when situating AI challenges to the contemporary human rights regime we can draw on the Legal Disruption Model. For our present purposes, I would like to highlight the prospect of legal disruption arising from different types of new affordances:

“uninvented affordances,” connoting the current lack of a technological capacity; (b) “unperceived affordances,” implying the lack of awareness of a new affordance; and (c) “unexploited affordances” which involve normative resistance against pursuing particular forms of behaviour which are otherwise possible and realisable.

Applying this aspect of the Legal Disruption Model, we can now see that the manifold challenges that AI applications will present to the contemporary human rights regime will, crucially, unfold and evolve over time. Thus, the human rights challenges introduced by AI applications that we are able to identify at present can, at most, constitute only a subset of the total range of issues that we face. Furthermore, because of the nature of these types of affordances, a veil prevents us from being able to identify the types of challenges that we may be confronted with in the future. A different way of approaching this problem would be through the concept of the “adjacent possible,” whereby we are only able to discern the proximate challenges but are unable to realistically and feasibly identify the character and impact of more distal developments.

The point here is that, even if we have a yardstick that is true and trustworthy in the contemporary human rights regime, the nature and character of the human rights challenges posed by AI applications will change over time according to the Legal Disruption Model, and if this is indeed the case, there will at minimum be a lag.
between the manifestation of a challenge and the mustering of an appropriate human rights response. At the other end of the spectrum, however, lie three further issues raised by the different types of affordances: that there will be multiple challenges to human rights presented simultaneously; that these challenges will be dynamic and “shape-shifting”; and that effective human rights responses and remedies may be elusive. As we will see in the following section, the contemporary human rights regime was designed retrospectively as a response to historical human wrongs. The exercise of recalibrating the contemporary human rights regime to face the challenges brought about by AI applications effectively seeks to deploy it against future and emerging challenges, revealing both its substantive and structural limitations in this regard.

The Contemporary Human Rights Regime

Drawing these threads together would suggest that the threats posed by AI applications are of a different character than those that the human rights system has been designed, and subsequently evolved, to respond against. In this section, I set out the contours and content of the contemporary human rights regime in order to more precisely identify its shortcomings. In doing so, the point is not to make a general critique of the regime as such: rather the purpose is instead to show the disjuncture between the nature and type of human wrongs that the system was intended to combat, and the issues raised by AI applications that may fall through the cracks of the system and remain unrecognized and therefore unremedied. There are six broad shortcomings, and the following sketch will necessarily be a caricature, but my aim here is indeed to accentuate the flaws that I see in the system.

First, the contemporary human rights regime is predicated upon the notion of discrete and enumerated rights. This is readily evident in the international legal documents, which stipulate for example the right to life. While such an approach may be effective in redressing particular historical wrongs, the outcome is a fragmented and disjointed system that interfaces with the complexity of the real world with difficulty. While a rejoinder may involve invoking a range of rights to cover a specific set of circumstances, often the wrong suffered may be greater than the sum of its parts that can be recognized by the siloed system of enumerated rights.

Second, human rights are founded upon the notion that these provide for individual protection and enjoyment. This is the premise that human wrongs are individually suffered, and therefore should be remedied on an individual basis. This premise suffers from a similar problem as the one above, because the impact of any real-world wrong must be filtered through its impact upon individuals, and again, the totality of the harm inflicted may be more than can be recognized through the wrongs that are inflicted on the individual.

Third, the rights that are recognized by the human rights regime impose corresponding duties imposed on the State. In essence, the State is therefore the only entity with the
obligation to protect and promote human rights, and therefore the only entity capable of violating them. This presents very familiar problems associated with non-State action, and raises rejoinders that the positive obligations imposed upon States cover some of this gap. But, even if that were the case, human rights still only protect against harms inflicted by the State or its agents and cannot shield against private action or omission as such.

Fourth, to invoke the machinery of human rights protection requires that thresholds of severity for infringement are met. Human rights protection mechanisms were designed to protect against infringements of a certain level of magnitude or severity, and were not intended to remedy just any slight. This is consistent with the notion that the recognition of a human rights violation carries significant weight and stigma, and therefore must be reserved for especially egregious actions.

Fifth, there are restrictions imposed by an implicit time-frame within which human rights violations must occur. Broadly speaking, this position implies that human rights protect against relatively clear-cut actions or discrete events at the cost of overlooking slow-burn or incremental processes. This prioritization of readily observable and time-bound events skews human rights protections towards certain types of activity, notably military and security operations, as well as legislative and policy acts.

Finally, human rights are capable of responding to certain types of normative restrictions. This is a subtle point drawing upon the existence of different regulatory modalities, and the prospect for technological management. As a sketch, normative modalities of regulation include law and social norms and set out incentives and disincentives for certain types of behavior but leave individual autonomy and choice largely intact. Non-normative modalities of regulation include architecture and code, and collapse undesirable behaviours into impossible actions thereby guaranteeing the desired outcome. The contemporary human rights regime, as a legally-based order, typically engages with normative restrictions to individual liberty and autonomy, but is largely blind to non-normative restrictions.

A different way of putting this point is that structural and systemic effects are imperceptible to the contemporary human rights regime. Sarah Schindler sets out the racially discriminatory impact of Robert Moses’ low-slung bridges on Long Island. The low clearance beneath these bridges readily permit access by private motorcars while simultaneously obstructing the passage of buses. This has become a paradigmatic example of the disparate impact of urban design because of the correlation between white people, affluence and private car ownership, and conversely of people of color with lower socioeconomic status and the use of public transportation. Features in the urban environment that exert differential accessibility therefore have strong correlational impacts upon social, and indeed, racial, groups. As Schindler puts it:

Instead of garnering support to pass a law banning poor people or people of color from the places in which he did not want them—which, if the intent were clear, would not be
permissible today—Moses used his power as an architect to make it physically difficult for certain individuals to reach the places from which he desired to exclude them.23

As Schindler notes, a clear body of subsequent jurisprudence rendered racial discrimination on the basis of law impermissible. Yet, despite legal and social progress, the racially discriminatory impact of the physical architecture remains to the present day beyond the reach and remedy of legal processes.

Drawing these threads together, we can make a rough sketch of the contemporary human rights regime. It is one where individuals enjoy specifically defined rights and freedoms against State interference, and one that will provide protection against severe infringements that occur in a discretely identifiable action or omission.

These parameters become problematic when they are considered within the framework developed by Scott Veitch that legal processes legitimize certain types of human suffering:

\[
\text{If the law can recognise some damages as injuries – if law sets some standard or test for ascertaining which damages will count as injuries – then the law also defines, by exclusion, that which will count as damage, but not injury. Such suffering is, legally speaking, legitimate.}^{24}
\]

Herein lies the crux for the contemporary human rights regime in the face of the challenges posed by AI applications. This problem is tautological: a human rights violation is what the human rights system recognizes is a human rights violation.

The corollary is that if the human rights regime fails to identify and recognize the challenges posed by AI, and this appears to be highly probable given the presumptions and limitations of the regime that I have just set out above, such shortcomings of the human rights regime will effectively legitimize the harm and suffering that flow from AI applications.

**Initial Shortcomings of the Human Rights Regime in Face of AI Challenges**

Having surveyed the landscape of the contemporary human rights regime, we are now in a position to identify some of its initial shortcomings in relation to AI challenges. This section will focus on the history and developmental trajectory of the human rights regime, and I will repeat this exercise again below in greater precision once we have modeled AI through different approaches, which will enable us to better identify the nature and character of the challenges ahead.

Contrary to the theory and rhetoric underpinning the modern human rights movement that emphasizes the universality of the legal catalogue of human rights and the inherent dignity of all human beings, a pragmatic approach suggests that human rights are historically contingent: they are the outcome of negotiations to remedy and redress past wrongs. Put differently, our contemporary catalogue of human rights constitute successful bids to recognize certain types of harms as severe injuries. Recalling Scott Veitch’s observation above, the modern human rights movement...
has precisely been about recategorizing certain types of harm under new and elevated category of injuries. The conclusion is that the contingency of the contemporary human rights regime does not necessarily reflect universal values. There are two prongs to this argument. First, John Danaher has demonstrated that human values have changed over time and space. At minimum, this should require a degree of humility with regards to our present axiological configuration, and, I would argue, militates against the possibility of asserting that the modern human rights catalogue should be preserved indefinitely into the future. In other words, human values are neither universal nor timeless, and it would be a mistake to ossify any set of such values into an eternal set of human rights. Second, as I have argued in earlier work, AI power exerts effects on human rights and human values at an intermediary level of power between discrete decision making and existential threats to humanity. In other words, as a powerful suite of general-purpose technologies, AI applications may exert pressure upon the very landscape of human values and social norms through which their effects will be understood, contextualized and contained.

The upshot is that, far from expressing universal, fundamental, and eternal human values, the historical contingency of the contemporary human rights regime and the evolving nature of human values demonstrate their susceptibility to external influences. Thus, a subtle challenge posed by AI is that its applications may tailor the very fabric of human rights and values in a way that is not readily perceptible. A different way of putting this issue is that a critical mass of AI-induced harm and suffering may need to be mobilized before the human rights regime can identify and engage with such new challenges. At minimum, to achieve this critical mass there would need to be a clear and consistent record of AI-induced human wrongs, and either a strong social movement or a captivating calamity to compel a human rights response. It is possible to construe this problem within the Collingridge dilemma between the information problem and the power problem. Before the widespread societal adoption of AI applications, their human-rights-relevant challenges cannot be readily discerned and hence cannot be easily guarded against (the information problem). Yet, after the widespread adoption of AI applications throughout society, human wrongs begin to manifest and to gain import, but by this stage it is very difficult to mobilize appropriate responses to contain and account for these harms. Thus it seems to be very difficult to prepare or to respond to the types of challenges that engage the Collingridge dilemma of control.

Finally, there is a clear risk inherent within the incremental and responsive approach implicit within the human rights regime’s retrospective orientation: that it will likely treat the symptoms rather than the root causes. In other words, human rights responses to AI challenges will be precisely that, responses. Absent proactive and holistic reform, the contemporary human rights regime will, at best, identify and redress specific manifestations of harm that have been thrown up by AI applications. The underlying generator of those harms, however,
will remain out of reach of the human rights system, because of the simple requirement that there need to be (potential) victims to move the system towards the identification of specific and concrete wrongs. A different way of putting this issue is that AI applications empower alternative ways for doing the same thing (better, faster, cheaper), which is itself a way of rephrasing that AI is a general-purpose technology. If this is true, then the underlying systemic changes wrought by the introduction of AI application remain beyond the reach of the band-aid responses that the human rights regime is capable of mustering.

Charting the Challenges Posed by Artificial Intelligence

Up until now, I have not sought to define Artificial Intelligence, nor have I offered a more sophisticated analysis of why AI might pose problems that the contemporary human rights regime will fail to protect against. So, why do AI challenges (fail to) fit into this picture?

A couple of preliminary points: I would suggest that we do not define what AI is and that we should not treat AI as the regulatory target.\(^{33}\) I do so because both of these approaches are brittle in the sense that small changes will render the regulatory regime obsolete or irrelevant. Attempting a definition is a lose-lose proposition: a rigorous definition of AI that comprehensively describes the technology or its functioning will be brittle against even the slightest subsequent change, and any regulation built upon this definition can be circumvented; a loose definition of AI, while more accommodating of change, will lack the precision necessary to underpin effective regulation. Treating AI as a regulatory target is also problematic because the regulatory problems are only partially influenced by the underlying technology.\(^{34}\) The difficulty can be seen in the “essential qualities” approach to regulation,\(^{35}\) which somewhat arbitrarily projects contemporary considerations as timeless problems inherent to the technology. “Salience” as the rejoinder notion, emphasized the context within which a technology is deployed to draw out relevant issues that may change over time.\(^{36}\) The concept of sociotechnical change compellingly foregrounds change in the sociotechnical landscape (itself a function of technology and the social, political, and economic context) as the appropriate locus for regulation.\(^{37}\) The overarching point that I would like to make is that we need to get beyond static and stable approaches to AI challenges for the purposes of framing human rights responses.

Instead, the question becomes how we should “understand” AI for the purposes of framing human rights responses? What are the models that we should use?\(^{38}\) What are the metaphors that we should deploy?\(^{39}\) Despite the oddity of these questions at first flush, they become critical when we shift the emphasis away from the nature and character of the technology itself, and instead towards the societal impact and human rights ramifications of AI applications.

An apt starting point is the ancient parable of the blind men...
and the elephant,\textsuperscript{40} which cautions against treating an elephant like a python simply because one has touched its trunk.\textsuperscript{41} Similarly, we should refrain from designing regulatory responses against AI applications based only upon a few plausible understandings of what AI might be and what it might do to us. Indeed, since the model and the metaphor of AI is the point of departure that leads to an array of path dependencies, it is critically important to work through the possibilities and their consequences thoroughly.

So what are the models that we use to understand AI? First, AI applications can be treated as a tool or instrument: here there is nothing new and the human rights regime can readily accommodate the problem with some tinkering.\textsuperscript{42} Yet to many scholars AI is ontologically different from mere technological tools\textsuperscript{43}: AI applications are placed within the liminal zone between agent and object based on the inference or projection of their independence, agency, and autonomy. This in turn befuddles the legal and moral concepts of control and concomitant responsibility,\textsuperscript{44} and curiously sets the stage for discussions as to whether AI applications should enjoy (human) rights protections.\textsuperscript{45} Thus, in this vein, AI applications are modeled as the substitutes for human beings and human activity — this is very much the flipside of treating AI applications as tools or instruments. These positions mark the ends of the spectrum ranging from treating AI applications as mere tools on the one hand to seeing them as attaining legal and moral agency on the other. They thus mark the polar opposites in terms of the corresponding human rights responses.

A broader model of AI applications is that these are powerful entities: they exert power over individual human beings or they exacerbate the power relations between human beings.\textsuperscript{46} Viewing AI through the lens of power permits the possibility to adopt and adapt contemporary human rights protections by refocussing them away from protecting individuals against the exercise of State power\textsuperscript{47} and facing AI power instead. If the human rights regime can be thought of as a means of holding the exercise of power to account, then it is “only” a matter of dogma and doctrine that the regime is directed solely toward State power.

Yet, as I have written elsewhere, AI applications may wield a different type of power than that which the human rights regime is attuned to, with the result that legal development would not be an adequate nor appropriate response.\textsuperscript{48} This view hints towards a different sort of problem that is not (only) legal in nature, and thus cannot have only legal responses. A different way of framing this issue is through the familiar model of control and responsibility: that human beings are “in,” “on,” or “out of” the loop. The type of relationship envisaged here simultaneously establishes and delimits the boundaries of moral and legal control and responsibility, but more importantly for our purposes, positions the human being as an agent in a position of power in this relationship by default. The implication is that human agency is dominant in “the loop” relationship, thus positioning AI applications as tools or instruments of human agency. Thus, this model is amenable to human rights oversight, insofar as the State
can be modeled as a social construction that is also an instrument of human power and agency.

The difference, however, with AI is that humans are increasingly “under the loop.” This is a subtle, but crucial point, because this implies the loss of human agency and control in the relationship with AI. Being “under the loop” means that human beings are transformed from agents to patients, from subjects to objects. This raises two different types of implications: first that AI applications can no longer be modeled only as instruments and thus are not necessarily within the realm of human control and responsibility; and second, that AI applications are not necessarily external to the human. This last point invokes Roger Brownsword’s notion of there being three generations of the regulatory environment and his view that “regulators would go beyond traditional normative signals and design of products and places by incorporating the regulatory design within regulatees themselves.” As we will see below in the context of invisible influences made possible by AI, it is not necessary to physically incorporate regulation within the regulatees as Brownsword originally implied.

There are two take-away points from this section. First, AI may exert or exacerbate different types of power that are not amenable to human rights oversight and response. The second is that, once regulation is incorporated into a regulatee, human rights protections are completely neutralized. This is so because, as mentioned above, the system of human rights presupposes the inherent dignity and autonomy of the individual human person who is the subject of protection and is premised upon external assaults to this dignity and autonomy. Where the source of human rights challenges has become internalized into the regulatee, the absence of an external threat nullifies the human rights regime.

Modeling AI to Systemically Exclude Human Rights Responses: A New Approach to AI and Human Rights

The orthodox approach to human rights is to identify the legal problem that a situation presents, thereby limiting the endeavor to identifying and resolving the legal lacunae that has been introduced. To caricature this approach, a real-world situation is dissected into various applicable human rights provisions, and then analysed to determine whether substantive human rights protections have been engaged, violated, and if so, whether their application is justified. In other words, the standard models, metaphors, and approaches operate to situate AI challenges as orthodox human rights problems. Thus, they present ways for us to frame or understand AI challenges within the language, scope and function of the human rights regime. Yet, such a presumption exerts a subtle selection pressure upon the range of possible models and sets the legal response trajectory towards legal development.

It should be clear from the argumentative flow of this paper so far that this position is neither necessarily justified nor conducive to achieving
the protective goals espoused by human rights. The foremost challenge, because of the ensuing path dependencies, is the curtailment of the full range of plausible ways in which to model AI and its applications. What if we sought to do the opposite? How might we model AI in ways that systematically exclude AI challenges from the human rights regime? Why might such an approach paradoxically reinforce human rights protections?

The orthodox approach to situating new human rights challenges has been to presuppose that the regime is applicable, and indeed the possibility that human rights might not apply is rarely, if ever, countenanced. Building upon David Kennedy’s observation that the human rights movement has crowded out other ways of understanding harm, I would argue that the human rights regime also monopolizes the arena of prospective harm and suffering. The upshot of this monopoly is that all harms — past, present and future — are deemed to fall within the ambit of the human rights regime, whether the regime is in fact capable of identifying, engaging with, addressing, or remedying those harms.

A different way of putting this is that human rights responses are, at most, confined to what Matthijs M. Maas and I have defined as Levels 0 and 1 of forms of governance responses which engage “business-as-usual governance” and “governance puzzle-solving.” This places human rights as a problem-solving mechanism, addressing problems that have already been identified and deemed to be important. Ordinarily, this is a sensible orientation to adopt, but it becomes maladaptive in situations involving significant socio-technical changes or legal disruptions where foundational presumptions have been significantly altered. Under such conditions, the potential problem space needs to be surveyed again for the simple reason that old problems may have morphed, and new problems may have emerged as a result of the seismic shifts.

We can ask these questions: what harms might AI applications threaten, that will not, perhaps cannot, be recognized as injuries, as human rights violations? Why might AI applications produce effects or outcomes that preclude human rights engagement or intervention?

My working hypothesis is that there is a residual core of the challenges posed by AI that are systematically excluded from the purview of human rights discourse because it falls entirely outside of models that the human rights system can engage with. Many of the behaviors, effects and outcomes of AI applications fall outside of the implicit foundations upon which the human rights system is built. Let me set out a caricature that might be called “the billiard ball model of human rights” where an external force perturbs an otherwise stable system. To do this, we adapt a familiar formulation from disaster studies:

Human rights violation

\[ \text{Human rights violation} = \text{Hazard} \times \text{Exposure} \times \text{Vulnerability} \]

If we treat a human rights violation as a non-mathematical function of “hazard,” “exposure,” and “vulnerability,” we can see the implicit assumption that AI is a hazard that is external to the system. Thus, AI applications are deemed to be an external threat to the individual enjoyment of human rights.
Two Key Pivots: Towards New Models for Understanding AI and Human Rights

What if we have this model all wrong! What if AI actually puts us in a new paradigm, while we are still deploying the concepts of human rights and human wrongs from a previous one?59

AI applications appear to be making two key pivots that are not accounted for by our orthodox approach to harms. The first pivot: We are used to thinking about AI applications making decisions about us, and that this is the core of the human rights challenges. Instead, AI applications interfere with our decision-making processes.60 While this is not a new point, it has far-reaching implications for human rights. If AI applications are no longer the external hazard that impinges upon our human rights and liberties, but become incorporated and embedded within us, this shift will effectively neutralize human rights protections, because the raison d’être of human rights is to protect the inherent dignity of the individual human being. Thus, the individual human being is the unit of protection: AI applications and their unrecognized threat, is that they embed themselves within, and interfere with, the agency of agents. As Shoshana Zuboff aptly reports on the goals of those who create and implement AI applications: “We can engineer the context around a particular behavior and force change that way … We are learning how to write the music, and then we let the music make them dance.”61 The desired behavior is incentivized, encouraged and enabled, but ultimately the behavior emerges from the individual agent(-regulatee) which is the locus of human rights protection. In other words, a human rights violation cannot be identified by a system that scans for external hazards because the challenge now lies within the victim, within the sphere of protection that is carved out by the human rights system.

The second pivot: We are used to thinking about an objective shared reality that is held and experienced in common. The ability of AI applications to tailor and to design personalized environments us towards atomized societies and isolated individuals. This leads to three important implications for human rights. First, if human rights are the result of successful bids to recognize and remedy past human wrongs, a fragmented society will necessarily undermine the revelation or emergence of new human rights protections. This is because a human wrong must be sufficiently widespread to impact a large number of people in a similar fashion, and it must do so in ways that are recognizable as collective wrongs and common injuries. One way to think about this is to look through the lens of key historical social movements, such as the women’s suffrage and the civil rights movements. One striking feature is the sheer mass of people participating in these movements, who were mobilized to right wrongs that were borne by large segments of the population. Where AI applications carve out filter bubbles and echo chambers, however, it
becomes difficult to identify common denominator grievances around which to build such movements for social change. Thus, if the harms that AI applications raise are not broadly shared, and are not shared in ways that are visible to others, then the momentum towards recognition falters, and the impetus towards recognizing new wrongs also dissipates.

Second, if individuals are isolated in their experiences, the absence of common benchmarks may render the very experience of “harm” imperceptible. Instead, such harm that may be suffered will become “just life.” The feeling of suffering a wrong is sterilized in the absence of shared suffering emanating from a common root cause because these wrongs have been converted into simply the individual’s life circumstances as a result of AI applications’ fragmenting of society. In a nutshell: human rights violations only have meaning in the context of a shared objective reality.

Third, it is perhaps trite to observe that context shapes behavior and that behavioral manipulation can occur in the absence of more overt instruments of law or political power. Langdon Winner famously wrote that there are two ways for artefacts to contain political properties:

First are instances in which the invention, design, or arrangement of a specific technological device or system becomes a way of settling an issue in a particular community … Second are cases of what can be called inherently political technologies, man-made systems that appear to require, or to be strongly compatible with, particular kinds of political relationships.62

The point here is one that I have advanced in more preliminary fashion elsewhere, when I suggested that AI exerted power at the level of human rights and human values.63 Instead of being concerned with AI power as such, the challenge posed by AI tailoring and designing personalized environments is that these engage with both of the observations advanced by Winner above. Thus, rather than legislating to pursue an objective or to secure an outcome, it is possible to design a system to settle that issue instead. And the implementation of particular AI applications may only be possible within certain political systems, and will serve to enforce and enhance that political system (the Chinese Social Credit System is a case in point).64 Situating this discussion back within the human rights context, Sarah Schindler has demonstrated the discriminatory impact of architecture and the way that the design of the built environment is treated more as “innocuous features rather than as exclusionary objects.”65

It is now possible to combine these observations concerning the political and discriminatory properties of artefacts and architecture with those that set out AI influences over our decision-making processes. The overall picture here is striking in that it is starkly different from those in which human rights might have a role to play. AI applications are not deployed as tools of the State, nor necessarily as instruments of power. Nor are AI applications making decisions about us that permit or prevent certain paths in life. Instead, the overall outcome here is that AI applications shape our internal subjective worldview, isolating us from a common shared reality. Our preferences are
amplified, our social networks carefully curated, our behavior sculpted by subtle shifts in choice architecture. Ultimately, because these effects shape who we are, while who we are is — at least at present — protected by the contemporary human rights regime, such phenomena fly below the radar.

From this analysis we can see that in addition to the various structural deficiencies of the contemporary human rights regime preventing it to contain the challenges posed by AI applications, perhaps the most insidious is how these applications become incorporated and intertwined with the human being and the notion of agency and autonomy that the human rights regime is intended to protect. But it is not only incorporation or interface with the human being that is the sole source of concern. While Roger Brownsword considered the incorporation of regulation directly into the regulatee as the third generation and final stage of regulatory travel,66 I would argue that the possibility to dynamically tailor an individual’s environment and context is integrally connected to the ability to control that individual. Put differently, what is the difference between influencing an individual’s behavior through incorporated or embedded forms of regulation, and influencing that same individual’s behavior by dynamically and adaptively altering the environment (and thus the structure of incentives and punishments) externally? In the first situation, it is the internal drive that is being tampered with; in the second scenario, the person becomes a puppet by virtue of being placed within a tailored straightjacket that is being controlled.

Concluding Thoughts

The upshot is that these two pivots identify situations where AI applications subvert the contemporary human rights regime by subterfuge. I started out by suggesting that there may be a tautology when examining human rights challenges presented by AI applications: that human rights issues manifest only when the human rights systems recognizes that there are human rights issues at play. If this is the case, AI does indeed present an array of familiar human rights problems (which I did not discuss here). Yet, I have also suggested that AI applications introduce subtle, yet systemic, distortions that lurk beneath the surface of untested presumptions, and that this will precipitate new types of wrongs. These wrongs will not only be invisible to the contemporary human rights regime, but they actually lock out human rights as a possible means of response. Ordinarily, locking out a single means of response might not be problematic, but recall that the human rights movement has effectively monopolized this space and that I have suggested that this might be true for the past and present, as well as for the future. This brittleness of the human rights regime leaves us in a lurch when it comes to the challenges posed and exacerbated by AI.

What I have hoped to do here is not to provide any proposals or answers to these challenges, but
rather to delve deeper into the nature and origin of the challenges themselves. Indeed, the deficiencies that I have identified in the contemporary human rights regime, when coupled with the multifaceted approaches to AI that we can plausibly take suggests a vast potential problem space. The two pivots in approach only scratch the surface as to the myriad ways in which AI applications can circumvent or subvert forms of monitoring and resistance (of which the contemporary human rights regime can be taken as just an example).

I hope that it is abundantly clear that we need to explore more of the potential problem-space before we settle down to solve the problems that we have identified. Problem-solving is a necessary component of the response, but it is only one aspect of it. How do we know that the problems we are trying to solve are indeed problems, or at least the most significant problems? And how might we have confidence that these are the main problems, if we do not look into the underexplored paths and spaces? I suggest that we need more proactive and systematic problem-finding work to figure out what we are not talking about. Furthermore, such problem-finding work must be constantly revisited and re-evaluated because the sociotechnical landscape is in constant flux. The most destabilizing aspects of AI in the context of human rights are not those issues that we have framed, explored and examined. Rather, it will be those unknown unknowns that will either blindside us unexpectedly, or which will be the source of a lingering, yet perpetual, sense that there are inadequacies and shortcomings in the existing human rights regime.

Drawing these observations and critiques together would suggest that a two-pronged complementary response would be necessary, each prong addressing a different facet of the challenge. The first would address the challenges posed by AI and its applications. In part because of the Collingridge dilemma, we cannot know the nature of the problems that AI will pose for human rights in the future, but the problems that AI is posing for us at present are already deeply entrenched and difficult and costly to change. At minimum, to address the challenges posed by AI, active, systematic, and continuous searches of the potential problem space need to be conducted both to attempt to meet the information problem and also to try to identify issues before they are compounded into power problems. These searches of the potential problem space must utilize broad ranges of metaphors and models in order to capture as diverse a range of understandings and implications arising from AI and its applications as possible.

The second prong complements a more varied and up-to-date understanding of AI impacts and implications with the development of more robust and resilient frameworks for human-rights-relevant responses (with an emphasis placed on the plural and complementary nature of these frameworks). While I have sought to dissect the structural limitations of the contemporary human rights regime in this paper, it is not only or perhaps even primarily, the internal principles and processes that raise the greatest problems for continued protection. Instead, it is the monopoly held by the contemporary human rights regime for responding to harm and suffering that makes responses simultaneously brittle and
prone to failure because there is only one barrier placed between hazard and failure. To be able to respond effectively to the challenges posed by AI to human rights, we are likely to need a multifaceted and dynamic set of principles and processes that are capable identifying, addressing and remedying the broad and evolving range of issues that emerge as AI applications are deployed in society.

Notes

[I would like to express my gratitude to Jonathan Jacobs and Margaret Smith for their kind invitation to speak at John Jay College of Criminal Justice of the City University of New York in March 2020 on which this paper is based.]

[Disclosure Statement: No potential conflict of interest was reported by the author.]

1 Two outstanding pieces set the broader scene for the human rights challenges posed by AI: Cohen, “Affording Fundamental Rights”; and Zuboff, “We Make Them Dance.”

2 My aim in this paper is to set out a guide and an overview of the inadequacies of the human rights regime when faced with AI applications. I have worked out some of the ideas in in previous work, including: Liu and Zawieska, “From Responsible Robotics”; Liu, “Digital Disruption”; Liu, “Power Structure of Artificial Intelligence”; Liu and Zawieska, “A New Human Rights Regime.”

3 David Kennedy suggests that “[a]s a dominant and fashionable vocabulary for thinking about emancipation, human rights crowds out other ways of understanding harm and recompense.” Dark Sides of Virtue, 9. When it comes to human rights law, the legal mechanisms that recognise damage, and which recategorise damage as harm and injury, play a pivotal role. On this point see Veitch, Law and Irresponsibility.

4 My use of “brittle” here is a heuristic for fragility, which is a concave sensitivity to stressors. This leads to a negative sensitivity to increase in volatility. See Taleb, Antifragile.

5 See Liu and Maas, “Solving for X?”

6 See ibid.

7 As the ECHR highlights in one of its rulings, “[t]he Court must also recall that the Convention is a living instrument which, as the Commission rightly stressed, must be interpreted in the light of present-day conditions.” Tyler v UK, No. 5856/72 (European Court of Human Rights 25 April 1978), paragraph 31.

8 See Letsas, “ECHR as a Living Instrument.”

9 See Bennett Moses, “Regulating in the Face of Sociotechnical Change.”

10 See Liu et al., “Artificial Intelligence and Legal Disruption.”

11 Don Norman explains that “[t]he term affordance refers to the relationship between a physical object and a person (or for that matter, any interacting agent, whether animal or human, or even machines and robots). An affordance is a relationship between the properties of an object and the capabilities of the agent that determine just how the object could possibly be used.” Design of Everyday Things, 11.


13 In Steven Johnson’s words, “[t]he adjacent possible is a kind of shadow future, hovering on the edges of the present state of things, a map of all the ways in which the present can reinvent itself.” Where Good Ideas Come From, 31.

14 This is further exacerbated by the nature of AI as a “general-purpose technology” capable of affecting all facets of
human activity. See Liu et al., “Artificial Intelligence and Legal Disruption.”

15 See for example, Clapham, *Human Rights Obligations*.

16 See for example, Osman v UK, No. 87/1997/871/1083 (European Court of Human Rights 28 October 1988).

17 The “minimum severity” standard is clearest in Article 3 for jurisprudence within the European Court of Human Rights. See Ireland v UK, No. 5310/71 (European Court of Human Rights 18 January 1978).

18 This time frame is different than the stipulated time-limits after the alleged human rights violation within which claims must be made.

19 See Lessig, “Law of the Horse.”

20 See Brownsword, “In the Year 2061.”


22 For structural forms of discrimination, see, for example, Schindler, “Architectural Exclusion”; Liu, “Three Types of Structural Discrimination.”


24 Veitch, *Law and Irresponsibility*, 86. (Emphasis added.) Thus, according to Veitch’s distinction, by definition “injury” are harms that are recognised by the law, while “damage” are harms to which the law is blind. This suggests that the initial categorization is critical, since responsibility need not be considered, let alone established, for “damage.”

25 As captured in, for example, the Preamble of the Universal Declaration of Human Rights (UDHR). UN General Assembly, “Universal Declaration of Human Rights,” UNGAR 217 A (III) (1948).

26 The fact that the modern human rights movement was born in the aftermath of the Second World War provides an example of this position. See Glendon, *A World Made New*.

27 See Dershowitz, *Rights from Wrongs*.

28 See ibid.


31 As David Collingridge puts it, “[w]hen change is easy, the need for it cannot be foreseen; when change is apparent, change has become expensive, difficult, and time consuming.” *Social Control of Technology*, 11.


33 But see, for example, Marchant, Allenby, and Herkert, *Growing Gap*.


36 See Balkin, “Path of Robotics Law.”

37 See Bennett Moses, “Regulating in Face of Sociotechnical Change.”

38 This question becomes pertinent if indeed “all models are wrong, but some are useful.” (A quip attributed to George Box.)

39 See Calo, “Robots as Legal Metaphors.”

40 See Saxe, *Blind Men and the Elephant*.

41 The metaphor itself may be overly restrictive: by failing to characterise the problem; by overlooking out-of-context problems; by missing out on the sociopolitical roots that shape research agendas; by missing out on the interrelation between AI and other potential problems; by adopting overly static rather than dynamic and adaptive perspectives; and adopting problem-identification from a human scale and perspective. See Liu and Maas, “Solving for X?”

42 Such thinking can follow the lines of “Legal Development” for example. See Liu et al., “Artificial Intelligence and Legal Disruption,” 233–42.

43 Here I am drawing an analogy with robotics. Ryan Calo has suggested that “robots, more so than other technology in our lives, have a social valence. They feel
different to us, more like living agents. The
effect is so systematic that a team of prominent
psychologists and engineers has argued for a
new ontological category for robots some-
where between object and agent.” “Robotics
and Lessons of Cyberlaw,” 532.

44 See Liu, “Refining Responsibility.”

45 There is a significant literature on this
point which I will not discuss, as it leads in
the opposite direction of the argument I
present here.

46 See Liu, “Power Structure of Artificial
Intelligence,” 209–14.

47 But see, Stammers, “Human Rights
and Power.”

48 See Liu, “The Power Structure of Arti-
ficial Intelligence,” 233–42.

49 Ibid., 209. Emphasis original.

50 See Brownsword, “In the Year 2061,”
30–1.

51 Ibid., 31. While any form of human-
machine interaction opens the door for
hybridization, the “reversal” heralded by
humans being “under the loop” transforms
them from being regulators to being regula-
tees. Thus, this reversal opens the door for
Brownsword’s third generation of regulat-
ory environments, as human beings
become regulatory patients.

52 See Liu, “Power Structure of Artificial
Intelligence,” 233–42.

53 Recall the above discussion surround-
ing the blind men and the elephant.

54 See Kennedy, Dark Sides of Virtue, 9.

55 Level 0 “involves an adherence to ‘normal’ governance processes, norms,
structures and concepts, however con-
structed. While it does not deny the emer-
gence of certain problems or challenges, it
can be slow to recognize them—and even
when it does, it will deny their fundamental
‘newness’. This governance level is there-
fore rigorously focused on solving prob-
lems within (or by extension or appli-
cation of) the existing governance
system.” Level 1 “still narrowly or primar-
ily emphasizes the importance of fixing
the direct problem at hand—that is, it
takes for granted a narrow rationale—but it
opens up for the possibility that doing
so may require innovation and (even far-
reaching) change in the regulatory tools or
governance processes.” See further, Liu
and Maas, “Solving for X?”

56 It thus stands in contradistinction to a
problem-finding orientation. See ibid.

57 See Liu, Lauta, and Maas, “Governing
Boring Apocalypses”; see also Wisner et al.,
At Risk; Perry, “What Is a Disaster?” This
formulation is only intended to express
the interrelationship of several variables.

58 By no means does this sketch limit
human rights challenges to only those
posed by AI, as it is intended as a general
model for understanding present and
future challenges to the regime.

59 I am here invoking Thomas Kuhn’s
idea of paradigm shifts. See Kuhn, Structure
of Scientific Revolutions.

60 See Yeung, “‘Hypernudge’”; Susser,
“Invisible Influence”; Susser, Roessler, and
Nissenbaum, “Technology, Autonomy, and
Manipulation”; Susser, Roessler, and Nis-
sembaum, “Online Manipulation.”

61 Zuboff, “‘We Make Them Dance,” 23.
(Emphasis in original).

62 Winner, “Do Artifacts Have Politics?,”
123.

63 See Liu, “Power Structure of Artificial
Intelligence,” 209–17.

64 See Mistreanu, “Life Inside China’s
Social Credit Laboratory”; Hvistendahl,
“Inside China’s Vast New Experiment.”

65 See Schindler, “Architectural Exclu-
sion,” 1954. See also the discussion above
with regards to the famous example of
Robert Moses and his low-slung bridges
on Long Island.

66 See Brownsword, “In the Year 2061,” 31.

67 See, for example, the “Swiss Cheese”
model of (human) error that places an
array of leaky barriers between hazards
and losses. Reason, “Human Error.”

68 Here I adopt James Reason’s terminol-
ogy. As I suggest above, the notion of human
rights challenges as external hazards may be
misleading. I would argue that all that
matters is that multiple and varied barriers
against failure are put in place.
Bibliography


