Conceptual Difficulties in the Transformation of Human Rights to the Realm of Artificial Intelligence

ANDRÁS HÁRS

Artificial intelligence has been seeping into various fields of international law for some time, affecting fields such as international humanitarian law – especially regarding the legality of autonomous weapon systems, but also intellectual property law and the legal profession as a whole. A conflicting zone encompassing many subfields is human rights, where an already sensitive subject that is open to debates and interpretation is met with rough questions. For instance, should and could human rights norms be transferred into pre-programmed entities? What relevance can human rights have to a non-human being that has been created, programmed and assembled by humans? Vast regional differences exist between the European, African and Inter-American systems with a lack of coherent structure in the Asia-Pacific region. Our understanding of human rights has also developed substantially over the decades, especially regarding norms on slavery, free speech, the prohibition of discrimination and the rights of women, of disabled persons and indigenous peoples to name a few examples. Furthermore, a vast array of international documents on human rights are political manifestos utilising expressions such as “respecting” and “ensuring” human rights as obligations for members of the international community. Since these provisions deliberately leave a lot of room for interpretation, it seems almost an impossible task to translate them to “binary code”, to a format that is digestible for an artificial entity. The article aims to answer these questions by analysing the abovementioned line of thought and combining it with various attempts at international regulation by states, international organisations as well as non-governmental organisations and think-tanks. The fundamental focus of this paper is to ascertain whether human rights and AI can be made compatible under the current framework of international law at today’s level of development.

Keywords: artificial intelligence, human rights, international regulation, transfer, compatibility

1 Senior Lecturer, Ludovika University of Public Service, Department of International Law, Faculty of Public Governance and International Studies, e-mail: hars.andras@uni-nke.hu
Introduction

A major achievement of the 20th century can be found in the adoption, codification and dissemination of human rights. This civilisational achievement stands for a guarantee so that individuals will not be treated as objects, no matter where they live or which societal group or construct they belong to, they will enjoy certain rights by virtue of being humans. The development of human rights norms is far from over, with many pertaining issues plaguing the system, ranging from lack of enforcement, regional differences, and criticism by authoritarian regimes, general disputes on the acceptance of new human rights, etc. Out of the current challenges faced by human rights norms, the dilemmas brought by the advance of artificial intelligence is a key concern for decision-makers. For simplicity’s sake, a working definition of human rights will be used as the following: unalienable rights that benefit each individual human being, foundational norms and achievements of the human civilizations which can be traced back to human dignity.2 The aforementioned definition merges elements from some of the leading scholars of international law who explain that the beneficiaries of human rights will always be humans. The obligation to respect human rights and refrain from violation in this context will affect the artificial intelligence (AI). However, since AI does not currently possess legal personality, other humans and abstract subjects of law are addressed such as companies responsible for development and dissemination of AI along with states whose role is to assure protection of humans from possible human rights violations through AI.3 The goal – and the easiest way to transfer human rights to AI and regulate AI-related conduct – shall remain in the hands of the states.

Therefore, the main aim of this article is to observe what major challenges may arise when we attempt to transfer human rights norms to artificial intelligence-based systems, and then to find solutions based on the current toolkit of international law. As a result, this article does not tackle the issue of ethical or moral questions nor programming aspects but merely the transference of human rights to artificial intelligence.4 It will not define what it means to be human, nor will it analyse transhumanist movements which would bring machine and human closer by integrating artificial systems into the human body.5

Difficulties of the current human rights regime

Contrary to how several decision-makers are referencing them, human rights are not a homogenous and well-defined set of norms that would be beyond and above debate

---

5 Bostrom 2005; Livingston–Risse 2019: 151–153. See also the attempts by Neuralink (https://neuralink.com/ approach/), or the prognosis by Ray Kurzweil (www.kurzweilai.net/the-transhuman-singularity).
regarding their content and extent. Therefore, a preliminary question arises before we can deliberate on harmonising the two fields: how are human rights affected or possible hindered by AI? The short answer – as referenced by several international organisations – lies in possible infringement by states and multinational companies resulting in mass surveillance, loss of privacy, adverse decisions without human control or interference, and generally, handing over decision-making to AI. Of course, in certain cases, reliance on AI can be beneficial as for instance instead of waiting for days or weeks for a simple application for official documents, AI can almost instantly “decide” the case and issue the requested certificates. However, in case of applying for a loan from a bank – with most banks using AI as of this moment to decide loan applications – if the process is not coupled with human oversight or the possibility to appeal to a stage where a human decides the fate of the application, the future of the individual will rest in the hand of an algorithm which can be quite problematic. As a result, making sure that certain human rights, especially the right to human dignity is observed, shall remain of paramount importance. However, putting this noble tenet to practice is not as easy as it sounds. In this chapter, the difficulties will be addressed, whereas potential answers by international organisations and possible theories in resolving them will constitute parts of the two following segments.

The first issue is what to transfer? As stated before, human rights are not a homogenous entity, and there are various catalogues of human rights exist which name, characterise and define human rights differently. Indeed, it is quite rare that an international treaty does all three. Which one of these documents do we accept as “the” catalogue or primary document? The International Bill of Rights (the 1948 Universal Declaration of Human Rights and the 1966 International Covenant on Civil and Political Rights and the International Covenant of Economic Social and Cultural Rights) has the benefit of having the largest legitimacy by virtue of being one of the earliest adopted, with the majority of the international community accepting (at least tacitly) their contents. It could also be argued that some international human rights treaties have since become ius cogens and therefore serve as the ultimate basis reflecting the values of the international community such as Convention against Torture and Other Cruel, Inhuman or Degrading Punishment or Treatment or the Convention on the Prevention and Punishment of the Crime of Genocide. Relatively newer international treaties also merit mentioning here, such as the 1989 New York

---

7 We have to bear in mind that such a software might not be AI or a vastly simplified version.
8 It needs to be taken into consideration that out of the 58 member states at the General Assembly of the United Nations, 48 voted in favour of adopting the Universal Declaration of Human Rights in 1948, while the ICCPR and the ICESCR have 173 and 171 state parties respectively according to the UN’s treaty database; UNTC 14531, 14668.
9 Or to be more precise, it can be argued that some of the content of international human rights treaties have since the time of adoption have not only become customary law but also peremptory norms of international law.
Convention on the Rights of the Child, as it seems to stir the least amount of debates and benefits from a wide, general acceptance from members of the international community.\(^{10}\) In contrast, the 1979 Convention on the Elimination of All Forms of Discrimination against Women lies at the forefront of debates both domestically and as a source of conflict between states either preferring the norms enshrined in the treaty or because of a special interpretation and practice of those norms.\(^{11}\) The abovementioned treaties merely represent directions the international community can take when deciding where to start but since there are dozens of human rights treaties on the universal level, choices are abound to begin with.

The second question is how to transfer human rights. Since human rights treaties are aiming for broad support, their provisions are formulated as to reflect a political compromise and operate with expressions such as “respect for human rights” is needed,\(^{12}\) that human rights need to be “observed”\(^{13}\) and “every step must be taken to make sure they are not violated”.\(^{14}\) From an international law standpoint, this is quite understandable as these constructs leave them open for various interpretations and incorporation into domestic rules as reflecting the dual regulatory nature of human rights: international norms manifest themselves as frameworks and directions, whereas domestic norms are defining their exact contents as per the other norms of the state in question. During the transfer of these norms to a binary equivalent, ambiguity has to be eradicated for the algorithm so as to provide clear results. If the treaty itself is not clear regarding its contents, if the contents can vary from state to state, then it can be argued that human rights norms exist on the level of principles, and coding principles might not be feasible at this stage of technological development.\(^{15}\)

On a very relevant side note, other sources of human rights also deserve to be mentioned here. Some of these norms exist as customary law – a set of rules derived from state practice and its accompanying opinio juris sive necessitatis – which is either not defined or parts of it take on the form of non-binding documents such as United Nations General Assembly resolutions. Since there is no hierarchy between international treaties and customary law provisions, it is extremely hard to balance treaty-based norms (which at least take on a written form where coding can start) with provisions that can be deducted from state practice and supported by judicial decisions. The process of proving the existence and applicability of a customary

\(^{10}\) Some peculiarities of the 1989 New York Convention will be addressed at later points as the situation is not as simple as described here.

\(^{11}\) See the “cultural war” debates in many countries and the strained foreign relations between several nations as a result of the implementation of women’s rights.


\(^{13}\) United Nations 1966a: Preamble.


\(^{15}\) This is only possible in science fiction, where advancements in technology have allowed for programming a human-like, complex brain with the capacity to understand a hierarchy of abstract concepts. See for instance the works of Isaac Asimov, where the so-called positronic brain represents the technological innovation allowing for the inclusion of binding principles.
law norm to a specific instance is the task of highly trained lawyers and such a process before a court can take years – something that as of this stage cannot be replicated artificially.16

The third question to be addressed is whether the instruments of international law can be incorporated into the programming of AI. If we take the example of the creation of treaties and the various stages of signature and ratification, the system of reservations and declarations which are abound in the context of human rights treaties – we see a complex web of connections which are currently deciphered by lawyers and courts leading to a highly complicated framework. This would not be a Gordian knot by itself, but some of the reservations attached to international treaties are hard to interpret and understand. A prime example can be observed regarding the 1989 Convention on the Rights of the Child. A seemingly simple and widely accepted international treaty, but several Muslim states have made a reservation to the part describing the freedom of thought and religion of the child.17 Among others, Afghanistan18 and Somalia19 have made reservations, stating that they will accept these provisions in accordance with sharia, the Muslim code of law. Without going into the debate on cultural relativism, from a purely structural standpoint, this brings a completely new set of norms into the equation. These are seemingly similar provisions concerning sharia by these two countries, but while Afghanistan applies sharia in all personal matters for non-Muslims as well including criminal matters, Somalia uses sharia solely for Muslims, and mostly for family matters and inheritance.20 Further complicating the issue is that these two countries belong to different schools advocating different understanding of sharia with Afghanistan belonging to the Hanafi and Jafari Schools and Somalia following Shafi’i teachings.21 Overall, if we add a system of law without which international human rights norms cannot be interpreted between many members of the international community to an already hard-to-define system, we arrive to a stage of almost insurmountable difficulties during the transference of these norms to an AI system.

Fourth, we arrive at the issue of regional differences. Human rights are layered, meaning that they not only operate on the level of universal treaties and domestic norms, but some continents have a well-developed regional human rights protection regime. The three most notable are the Council of Europe, the African Union and the Organization of American States. These have their own regional treaties, court

---

18 “The Government of the Republic of Afghanistan reserves the right to express, upon ratifying the Convention, reservations on all provisions of the Convention that are incompatible with the laws of Islamic Shari’a and the local legislation in effect.”
19 “The Federal Republic of Somalia does not consider itself bound by Articles 14, 20, 21 of the above stated Convention and any other provisions of the Convention contrary to the General Principles of Islamic Sharia.”
21 See also Coulson 1994.
system and judicial practice creating a unique set of norms. They possess their own understanding of human rights, and in some cases push states towards heightened protection, whereas other continents lacking such a regional cooperation or where the regional organisation does not have a court system (such as Asia, the Middle-East, Australia, and the Pacific) are missing this layer of protection entirely. Should AI-based technologies reflect this regional difference, and if so in what way?

Last but not least, it is widely understood that the development of human rights is not over. New treaties are being adopted, customary norms are forming, and judicial practice is changing. Human rights are a relatively new phenomenon, deriving from the 17th and 18th centuries, gaining momentum during the American War of Independence and the French Revolution of 1789. Since then, major developments have occurred, such as the abolition of slavery worldwide, the gradual reduction of discrimination of women and the protection of societal groups (minorities, indigenous people, people with disabilities, etc.) among others. Progress does not stop here. Society can agree on new norms, and previously achieved consensus can be called into doubt and reopened as societal and political dialogue providing lawyers and decision-makers with a plethora of tasks. Is it possible from a technological standpoint to incorporate development and furthermore, could we and should we outsource this task to an AI system?

**Current regulatory attempts**

Concerns of states and individuals have not gone unnoticed by international organisations, and several of them, namely the European Union (EU), the United Nations Educational Scientific and Cultural Organization (UNESCO), the Organization for Economic Cooperation and Development (OECD), as well as the Council of Europe (CoE) have begun harmonising the conduct of states and proposed regulations that significantly affects human rights.

One of the earliest regulators was the OECD, which, in its Recommendation of the Council on Artificial Intelligence in May 2019 has emphasised the importance of a “trustworthy and responsible” approach towards AI by states. The document’s approach is based on “respect [for] human rights and democratic values” deriving from the Universal Declaration of Human Rights and the United Nations’ Sustainable Development Goals. Highlighting the importance of a human-centric development, the OECD has also promoted a toolkit for states who want domestic regulation and at the same time, the recommendation aspires to form the basis for all future negotiations.

2021 was a particularly fruitful year in terms of international regulatory attempts as in April, the EU has discussed the first version of its AI Act, followed by the

---

22 Çali et al. 2018: 130.
23 See the re-igniting debate on abortion in the United States for instance.
24 OECD 2019: Art. 1.2. a).
UNESCO recommendation on ethics and principles in November, and the year ended with the CoE ad hoc working group’s outcome document in December. The EU’s AI Act is somewhat unique in that once it gets adopted, it will be the first binding source of law for EU Member States when it comes to AI regulation with a great effect on states attempting to have economic and financial deals with the European common market.\textsuperscript{25} The EU is attempting to create a regulation that is human-centric, safe, and transparent, possesses human oversight and has adequate tools to mitigate risks arising from the use of AI-systems. Its human rights approach is based on the European Charter of Human rights and it requires norms to be “in accordance with EU values, fundamental rights and principles”. A novelty can be found in the classification of AI systems based on the risks they posed in violating human rights and it is exactly in these “high-risk” systems that the EU takes an adamant position concerning its regulation.\textsuperscript{26}

Compared to the EU’s robust approach, UNESCO is a bit more generalised in nature, albeit with 193 member states, the organisation is aiming at a large-scale consensus which can only be achieved through significant compromise. UNESCO’s recommendation references the largest array of human rights norms, such as the UN Charter, a wide scope of international treaties adopted under the aegis of the United Nations and the work of universal institutions like the UN’s Human Rights Council. It relies on the central tenet of human dignity and the freedom from harm and subjugation – echoing strong societal fears of losing control of AI systems. Its uniqueness can be found in the policy actions it presents to states attempting further harmonisation, seeking good practices or a strong domestic regulation.\textsuperscript{27}

Arriving in late 2021, the outcome document of the Ad Hoc Committee on Artificial Intelligence (CAHAI) is peculiar because it paves the way for a regional human rights treaty to be adopted in the following years specifically addressing the human rights concerns of AI-based systems. It echoes UNESCO’s reliance on human dignity as a foundational norm and the EU’s approach regarding risk assessment along with the OECD’s reference of the Universal Declaration of Human Rights. It connects human rights with the concepts of democracy and the rule of law, and presents them as inseparable tenets that can only be realised together. It is unclear, in which way the CoE will develop these notions or when will a new international treaty be created, but even in its current form, the CoE outcome document shows the interest of member states in a strong, human rights-based cooperation for the years to come.\textsuperscript{28}

Besides international organisations, states have also been active in adopting national AI policies – close to 200 by 2022 – which handle the question of human rights very differently to one another. Depending on the state’s approach, economic, scientific and research potential along with the measured or perceived attitude of

\textsuperscript{25} Franke 2019: 4.
\textsuperscript{26} European Commission 2021: Art. 2.2, 2.4, 3.3.
\textsuperscript{27} UNESCO 2021: Art. 8, 50–52.
\textsuperscript{28} CAHAI 2019: Art. 4, 7, 16, 19, 21.
voters, a great degree of variety can be observed. Overall, it can be stated that human dignity appears to be the cornerstone of all regulatory attempts, which nonetheless either remain non-binding and somewhat vague (OECD, UNESCO) or currently under debate and deliberation and prone to change before taking on their final forms.

Towards possible solutions

With the currently existing (and proposed) international regulations providing little help in answering the questions raised in the previous chapters, it falls to science to untangle the web. This part aims at finding common ground between the international framework of human rights and the methods used by artificial intelligence.

The first two questions should be merged and extrapolated in one segment in order to better understand the proposed solutions: what and how to transfer? It is safe to say that at our current level of technological development, transferring the entirety of human rights material to AI would be an impossible task. Therefore, it should be seen, whether a certain portion of the human rights law can be transferred. International treaties – universal or regional ones in particular – are a good place to start as their text is certain, finalised, and not likely to change. They also benefit from larger legitimacy deriving from broad support by members of the international community. As a result, treaty texts can be used as a primary document which serves as the initiation point for later reference. Commentaries and academic articles can be used by language analysis AI for the algorithm to better define the meaning of statements and concepts found within the treaties. One of the solutions is to start small, select an international treaty which has a judicial mechanism such as the European Convention on Human Rights and extensive judicial practice such as that of the European Court of Human Rights, and use deep learning to predict human rights violations. So far, this is nothing new as similar systems have been used to predict court decisions. The novelty lies in the modular developments that can be applied. For instance, joining this system with national databases could predict human rights violations on the domestic level, resulting in dozens of sub-databases. Creating a similar system to other regional levels and compare the level of human rights protection would also be possible almost instantaneously. As a following step, deep learning or reinforcement learning can be used in a controlled environment (commentaries, interpretations, state practice) to advance the understanding of human rights by the AI. This would serve a dual purpose: new state policies and detailed descriptions of new AI systems could be run on this AI which would analyse the document, then spot and signal possible human rights violations, so that these could be avoided. As can be seen, the

29 This has been widely analysed with a major study examining the growing domestic policies almost every year. See also Field et al. 2020; Fukuda-Parr–Gibbons 2021: 33.
Conceptual Difficulties in the Transformation of Human Rights…

initial model would be very limited in its applicability, as it would have to be narrowed down along the lines of international human rights law to only binding norms that have a background in judicial practice, such as treaties, and this would severely limit the application of the model, but it would nevertheless be workable.

Difficulties arise when we attempt to use the method above a treaty without a judicial mechanism, as it would not have a database that reflects practice, and it is doubtful whether special constructs which do not work as judicial mechanisms – such as the universal periodic review of human rights by the United Nations31 – would work.

The third question was for the possibility of the toolkit of international law be transferred to an AI system. In this respect, AI is a highly adaptable system which can learn the hierarchy of norms easily.32 If we teach it the meaning of peremptory norms through treaty texts, state practice, court decisions and the “meaning” of certain terms such as torture or genocide, it could use those terms quite well. It could even apply a priority scale to differentiate between treaty-based obligations and soft law. The same is true for declarations and reservations concerning international treaties if they apply to a certain, well-defined section of the treaty text. Using broader terms, such as applying the treaty through the lens of sharia as the example in previous chapters, would prove to be an insurmountable challenge to the system. It is also unclear if competing norms or those that can collide, for instance the freedom of speech and the freedom of religion or personal self-determination, and the rights to life could be resolved, even if it has some judicial practice.

The question of having regional systems also merits some reflection. Since regional human rights systems have effective judicial control mechanisms, they are to be given priority. As there is no hierarchy between universal and regional human rights, with the universal level lacking a court structure, this is not a problem. Nor is there any overlap between regional systems, as states belonging to one of the three major systems (Africa, America, and Europe) are not members of the other.

Lastly, the possibility for societal development to be incorporated deserves to be tackled. Human rights are a developing concept, and with the advancement of human civilization, our understanding of what it means to be human and how humans are to be protected changes also.33 We do not even know where the future of human rights lies, as for instance the debate on abortion and in general the right to life was once again brought to the forefront of attention as the Supreme Court of the United States changed the 50-year consensus established in Roe v. Wade.34 Theoretically, it is possible to predict and guide the development of human rights as in economics and

---

31 Universal Periodic Review: www.ohchr.org/en/hr-bodies/upr/upr-main
32 Akin to Isaac Asimov’s three (later four) laws of robotics.
marketing, predictive models and algorithms are abound. In due course, a similar system operating on a much larger scale can be envisioned. However, if we wish decisions to be left in the hands of mankind, this should not be made possible, as the discussion to guide the direction of human rights development should remain in the hands of human and not “outsourced” to AI.

Concluding remarks

When it comes to international organisations, the way they understand AI and human rights has two sides: they either see the risks and possible human rights violations or they observe that in general, the use of AI can benefit humans, provided that certain safeguards and accountability mechanisms are put in place. The human rights provisions found in these recommendations and draft documents are vague, relying on either principles, fundamental concepts such as human dignity or calling for a “respect of human rights” without elaborating on the tasks of the international community and its members. Naturally, it can be assumed that decisions in this context will be made by states on the domestic level, and international law has little part to play.

This article took on a different approach, and tried to illustrate what are the exact problems and dilemmas when we attempt to use two systems that were designed independently, decades (or centuries) apart by people with vastly different mindset and for different problems. In this paper it is argued that some of the concerns can be resolved by AI-based technologies even at the current level of development such as the creation of an algorithm predicting human rights violations based on an international treaty, and its accompanying judicial practice that can be modularly improved to handle domestic practice as well. A much larger scale system could also be developed in the future, as AI can handle many aspects of international law such as the court structure, reservations and declarations, and the hierarchy of norms. Other parts of public international law would be very hard to fit into this system such as customary international law, competing opinions in science and the relation between “soft” and ‘hard’ law. It can also be stated that some questions such as the directions of societal and the ensuing legal development could theoretically be predicted and outsourced to AI, but it is not in our interest as humans to lose control in such a manner. In essence, the proposal enshrined here is to use AI in order to predict human rights violations by AI, and through modular development, it can be expanded in many directions. Nonetheless, this is merely a thought experiment which was created to form the theoretical basis for future research. The next step lies in the creation of such a model, attaching the first extensions, and then testing it on a domestic policy or an AI system’s detailed description, so that violations can be spotted, signalled and then prevented.
References


Conceptual Difficulties in the Transformation of Human Rights...


Case