JOURNAL OF CRIMINOLOGY AND CRIMINAL JUSTICE

Vol.2, No.1 2026 E-ISSN. 3137-0543. P-ISSN. 3137-0535 OPEN ACCESS

RESEARCH ARTICLE // DOUBLE-BLIND PEER REVIEWED

Artificial Intelligence and Human Rights: Balancing Technological Innovation and Fundamental Rights Protection

Ali Anwaar[1] **Buryzhan Yessov**[2]

[1] Nazarbayev University Graduate School of Public Policy (NU GSPP) , Astana, Kazakhstan. E-Mail: ali.anwaar@nu.edu.kz
ORCID: https://orcid.org/0009-0004-6727-0184
[2] Graduate in Political Science and International Relations Nazarbayev University Graduate School of Public Policy (NU GSPP) , Astana, Kazakhstan. E-Mail:bauyrzhan.yessov@nu.edu.kz

ABSTRACT

Artificial intelligence has moved rather quickly from a specialist debate into ordinary institutions: police forces, border systems, hiring desks, clinics, social media and consumer markets. It is this move that does not fit the picture of a rights violation I was taught, and why I started this paper. The standard image is a decision by a named official or a company exercising a public function that can be challenged. Often there is an AI system in between. It is trained on old data, it may be opaque even to its designers, and it can be deployed across borders by a combination of public bodies and private firms. This article considers the effect of that change on privacy, non-discrimination, freedom of expression and the right to a remedy. It is a doctrinal and policy-based method. I studied the main human rights instruments, a small set of well-documented cases and investigations, and the newer regulatory texts, especially the EU Artificial Intelligence Act, the GDPR, the Council of Europe's 2024 Framework Convention, and the UN Guiding Principles on Business and Human Rights. I am not arguing that human rights law has nothing to offer. Law is built for identifiable actors exercising judgement. Statistically mediated systems expose the gaps in that design. The recommendations at the end are correspondingly practical: impact assessment that is not self-certified, remedy that a person can actually reach, and due diligence duties that attach to private developers as well as to states.

01 ARTICLE HISTORY

Received	11-08-2026
Accepted	03-10-2026
Published	10-10-2026

02 KEYWORDS

Artificial intelligence; human rights; privacy; non-discrimination; freedom of expression; algorithmic accountability; international law

03 CORRESPONDING AUTHOR

Ali Anwar
Ali.anwar@nu.edu.kz

CC BY-NC 4.0 © 2026 The Author(s). Published by JuriSocial Research Center Private Limited. This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International Licence (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial use, distribution and reproduction in any medium, provided the original work is properly cited.

https://doi.org/10.68740/jjccj23

1. Introduction

I keep returning to one simple mismatch. International human rights law, developed in the aftermath of the Second World War and elaborated through treaties, regional courts and custom, presumes a specific type of infringement. You may be named. The conduct is attributable. The court or commission may examine the file. Artificial intelligence doesn't always refuse that model, but it distorts it. An alert from facial recognition technology can send police after the wrong person. A hiring tool can silently downgrade applications that mention a women's organization. A recommendation system can push inflammatory material further than any editor would have wanted to push it. In each case the outcome has a rights dimension, and in each case the way back from outcome to responsible actor is more difficult than the textbook suggests.

They are no longer hypotheses, and that is why I was not comfortable leaving the paper at the level of principle. In Bridges v Chief Constable of South Wales Police [2020] EWCA Civ 1058, the Court of Appeal found that the use of live facial recognition by South Wales Police was a breach of Article 8 of the European Convention on Human Rights, as the legal framework did not set out sufficiently clearly who could be placed on a watchlist or where the cameras could be deployed and was also a breach of the public sector equality duty, as the force had not taken reasonable steps to investigate whether the software was biased on the basis of race or sex (Bridges v Chief Constable of South Wales Police, 2020). I read that judgement as a privacy case as well as an equality case. It did not prohibit the technology. It said the force had failed to do what the existing law required it to do.

Clearview AI is the commercial version of that same problem. The company scraped publicly available photos and built a searchable biometric database without the consent of the people in the images. In 2022 the Italian, Greek and French data protection authorities imposed €20 million in fines each; the UK Information Commissioner's Office imposed £7.5 million in 2022; and the Dutch authority imposed a fine of €30.5 million in 2024 (CNIL, 2022; Garante per la Protezione dei Dati Personali, 2022; Hellenic Data Protection Authority, 2022; Information Commissioner's Office, 2022; Autoriteit Persoonsgegevens, 2024). What struck me reading those decisions together is not the size of any one fine. It is that a firm without a European establishment has been able to view a pile of sanctions as a cost of argument rather than a stop order. Collection is contested. That's a remedy problem, not just a privacy problem,

The employment example is older. But still relevant. In 2018, Reuters reported that Amazon had abandoned an experimental recruiting tool after it was discovered that the system, trained on a decade of mostly male technical hiring, had learned to downgrade CVs containing the word "women's", as in "women's chess club captain" (Dastin, 2018). Amazon said the tool was never used in isolation to make hiring decisions. I make that caveat because the story is sometimes told as if the company ran all its recruitment through a biased machine. The narrower fact is already sufficient. A system can reproduce a discriminatory pattern without anyone typing in a discriminatory rule.

The structure is the same but Cambridge Analytica is in another rights box. "Personal data collected from Facebook users, without sufficient consent, was used to build profiles for political advertising in the run-up to the 2016 US election and the UK's European Union referendum (Cadwalladr & Graham-Harrison, 2018). First thing was privacy. If you take the reporting at face value, the extra damage was to informed participation in public life. I am loath to say micro-targeting "decided" either vote. The evidence does not require that claim. The weaker claim, that profiling can be used to shape what people see, is needed. That freedom of expression includes a right to seek and receive information, not merely a right to speak.

All these episodes have a feature that I want to keep in view. An AI system operated by a state body or a company or both produced an outcome with implications for privacy, non-discrimination or political participation, and the mechanisms for preventing or repairing the harm were slow, partial or absent at the time. The paper poses three questions. How do design and deployment affect privacy, equality, expression

and remedy? What features of AI – data dependence, opacity, cross-border applicability and scale – explain the gap between harm and remedy? How are international, regional and national frameworks responding?

The order is a simple one. 2. Literature review We describe the method in Section 3. 4. Applications We describe the applications in this section. Privacy, non-discrimination, expression and remedy are each considered in turn in sections 5-8. Section 9 compares the legal reactions. Section 10 pulls together the threads. Recommendations are presented in Section 11. Section 12 concludes. I have not tried to write as if every gap were a failure of imagination. Some of the gaps are built into a system of responsibility that still seeks a person who decided.

2. Related literature

2.1 Human rights architecture and AI

A growing body of literature seeks to map AI onto the existing catalogue of rights: the Universal Declaration, the International Covenant on Civil and Political Rights, the International Covenant on Economic, Social and Cultural Rights and regional texts such as the European Convention. The usual conclusion, and one I share, is that the norms are technologically neutral enough to apply. Privacy, non-discrimination and expression don't expire because the means of interference is a model, not a memo." It does not follow that the norms are already in operation. There remains work to be done, both interpretive and institutional.

The doctrine that has mattered most to me in this reading is positive obligations. The European Convention requires that the state not only refrain from harming people itself but also sometimes protect people from other people. That doctrine is quietly working hard in arguments that governments should regulate private AI even when the state is not the user. I think the principle of the argument is good. I think it is also easy to overplay. A positive obligation is an obligation to take reasonable steps not an obligation to prevent every bad outcome a platform is capable of producing.

2.2 Business and human rights

The second literature starts from the UN Guiding Principles on Business and Human Rights, adopted by the Human Rights Council in 2011. The responsibility of business to respect rights is distinct from the duty of the state to protect. The responsibility of business includes due diligence: identify, prevent, mitigate and account for adverse impacts connected to their activities (UN Human Rights Council, 2011). The OHCHR B-Tech project has sought to do this for the tech sector, including design and deployment of AI. I find the translation useful but lacking. Useful, because much of the relevant development is private, and cross-border, so a state-only framework will miss the developer. Incomplete because Guiding Principles are not a treaty. A non-enforceable responsibility looks like a policy preference from the outside.

This is why I am more interested in such later mandatory due diligence statutes in Europe than another set of principles. They do not decide jurisdiction. They do shift the discussion from "should a company look" to "a company must look, and demonstrate that it looked." Whether that model can be extended to AI developers without becoming a box-ticking exercise is one of the open questions in Section 11.

2.3 Digital constitutionalism

A third strand, sometimes classified under the rubric of "digital constitutionalism," investigates the ways to embed constitutional values in the governance of private digital power (Suzor, 2019; Nemitz, 2018). The analogy is that big platforms have a kind of power over speech and visibility that is functionally similar to state power, even if the legal categories are different. Klonick (2018) called platforms "new governors." Douek (2021) argues that speech governance online can't claim every post is a trump, and that probability and proportionality better describe what moderation actually is. I'm sold on the description. I am less convinced that the analogy results in a simple legal transplant. A company isn't a country. As I read the literature, the point is that the power is real enough to need constraint, not that every public-law doctrine should be copied across.

2.4 What I extract from the three strands

Taken together, the three strands suggest that neither classic state responsibility nor voluntary business standards will do the job on their own. An integrated approach would use human rights norms to express what must not happen, business and human rights tools to influence private developers, and binding regional or domestic regulation to give life to duties. This is the approach this article attempts to adopt, limited to privacy, non-discrimination, expression and remedy. I make no claim that the integration is elegant. It is still uneven in the instruments I discuss later.

3. Methodology

Method: Qualitative, doctrinal and policy-analytical. I used three types of material. First, human rights instruments and the cases applying them, particularly S and Marper v United Kingdom in the European Court of Human Rights and Bridges in the Court of Appeal. Second, episodes documented in litigation, regulatory decisions or sustained reporting: live facial recognition, Clearview, the Amazon experiment, Cambridge Analytica, and the Myanmar investigations. Thirdly, the existing regulatory texts: the GDPR, the EU Artificial Intelligence Act, the Council of Europe Framework Convention, the UN Guiding Principles and B-Tech guidance, and the UNESCO Recommendation.

These episodes were chosen because there is enough in the public record to argue from. That is a drawback too. There are harms that never come to judgement or the newspaper, perhaps the larger set. The reading is theme based. The same difficulties I found across domains: attribution, cross-border deployment, and the gap between a formal right and a usable remedy. This will not output a percentage. It shows if the legal architecture is fit for the purpose it is asked to govern.

Another limit I should state rather than bury is time. The AI Act was adopted in June 2024 and is still being phased into application. The Council of Europe convention was opened for signature in Vilnius on 5 September 2024 and has not yet entered into force (Council of Europe, 2024). Any comment I make on their practical effect is a reading of the text, not a practice evaluation.

4. AI Applications that Impact Fundamental Rights

4.1 Biometric surveillance

Facial recognition redefines what anonymity can mean in a public space. Live systems scan crowds and compare faces to a watch list. And that's what South Wales Police were doing at public events and that's what Bridges looked at. Commercial systems have a similar effect but differ in ownership. Clearview's product was a search tool sold to clients, built from images of people who had not consented to be enrolled Even if no officer stops you that day, if a face can be looked up at scale, then the practical content of "I was only walking down the street" changes.

I have also taken note of the technical literature on error rates, since this is what the equality discussion is based on. Buolamwini and Gebru (2018) discovered that commercial gender-classification systems were more error-prone on darker-skinned women than on lighter-skinned men, attributing that gap mostly to the composition of the training and benchmark data. Demographic differences in false match rates were also reported later by NIST in vendor tests (Grother et al., 2019). These studies do not prove that all police deployments are biased. They do make it unreasonable for a public body to deploy a system and not ask.

4.2 Employment, credit, and criminal justice

A second cluster is decision support: hiring, credit, and tools that estimate risk in the criminal justice system. The Amazon episode is the clearest public example in employment because the company itself described the failure. Credit scoring is similar , with less public documentation . Tools for recidivism, discussed in the fairness literature and in investigative reporting on systems such as COMPAS, bring the issue back again in a setting where liberty is at stake (Angwin et al., 2016). These examples are to show a problem with training data, not to say that all models in these fields are discriminatory. The legal point is more narrowly. If a system is trained on an unequal history, equality law cannot stop at the absence of a forbidden variable in the code.

4.3 Moderation and recommendation systems

Platforms use models to moderate and rank speech now because there's no way a human can review it all. That makes two mistakes at once. Removing lawful speech, including satire, counter-speech that a classifier may not recognise as such. Illicit or dangerous speech is left up. Third problem is recommendation systems. They are usually optimised for engagement, rather than correctness. UN investigators looking into the violence against the Rohingya found that Facebook has been a useful tool for hate spreaders, and the company later admitted that it had not done enough (UN Independent International Fact-Finding Mission on Myanmar, 2018). I do not take that report as a finding that an algorithm "caused" the violence. I take that as evidence that amplification is not a neutral pipe, and that the right to life and security can sit in the back of a speech-governance debate.

4.4 Profiling and micro-targeting

Cambridge Analytica made this real for a non-specialist audience. Facebook used data from users to build profiles for political ads without proper consent (Cadwalladr & Graham-Harrison, 2018) The cleanest legal angle is a violation of privacy. The democratic worry is harder to show, but it is worth mentioning. If freedom of expression is about seeking and receiving information, then a system that sorts political messages according to their predicted susceptibility is more than a marketing technique. It is an intervention into the conditions of public debate. The size of the intervention in 2016 is still debated. Whether or not that dispute is settled doesn't affect the legal category.

4.5 Borders, welfare and health

The abstract of this paper references border control and healthcare and I didn't want those to remain as a list. Automated risk scoring and biometric matching at the border determine who is questioned, delayed

or refused entry. The affected person is often a non-citizen, with less procedural rights than a resident would have in a criminal court and little time to challenge the score. I've not seen a judgement that does for borders what Bridges did for a British high street. That's part of the problem. A system can be rights-relevant and still not be one of the cases that scholars usually cite.

The pattern for welfare and health care is more muted. Surveillance looks like a model for ranking benefit claims or flagging a patient as low priority. Looks like administration. The stakes are still familiar: equality, if the error is on a protected group; privacy, if the file is constructed from data the person did not expect to be pooled; and remedy, if the decision is a score with no reasons. I'm not saying every clinic has a documented scandal like Clearview. I say that the change from a police camera to a benefits screen makes no difference to the legal issues. What you see is what changes. A wrongful arrest is in the headlines. A wrong de-prioritisation often does not.

5. Right to Privacy

It's not without reason that privacy is the right people go for first. Article 12 of the Universal Declaration , Article 17 of the ICCPR , Article 8 of the European Convention , all deal with it in different language . Most machine-learning systems require data. Some others are to identify people. It's a straight fit.

S and Marper v United Kingdom is still the case I would start a student off with. The European Court held that it was not fair to retain DNA profiles and fingerprints of persons arrested but not convicted en bloc and this was a violation of Article 8 (S and Marper v United Kingdom, 2008). This was before live facial recognition came along. Bridges used it anyway, and I think he was right to do so. The transferable point is not 'biometrics are unlawful'. The point is that retention and matching need a legal basis that is clear and foreseeable, and a balance that has actually been struck. The South Wales Police failed that test, as the policy was not sufficiently constraining on who could be placed on the watchlist or where the system could be used (Bridges v Chief Constable of South Wales Police, 2020).

Clearview demonstrates the limits of taking a police force that is inside a jurisdiction and moving to a company that scrapes the open web. The GDPR's requirements on lawful basis, purpose limitation and data minimisation apply in principle to training sets and to biometric search. European authorities have said as much, in decisions that also rely on the application of the Regulation outside establishments in the Union. In writing, enforcement has been strong; in collection, weak. The Dutch fine in 2024 suggests that the legal theory has not been abandoned. It also suggests that a fine is not a deletion order that the rest of the world will execute (Autoriteit Persoonsgegevens, 2024).

Once a region has a data protection statute, there is a temptation to think that privacy is solved. I don't think the record supports that. The statute tells us the questions, what is the basis, what is the purpose, how little data will do. It does not in itself produce an investigator who can view a training set, or a court order that applies to a firm with no local assets. That is the bridge from this section to the remedy section.

6. Non-Discrimination and Equality of Rights

Equality provisions, such as Article 7 of the Universal Declaration, Article 26 of the ICCPR, Article 14 of the European Convention, and domestic statutes such as the UK Equality Act 2010, are triggered when outcomes systematically differ by race, sex and other protected grounds, even if the model never receives those grounds as an input. It is the Amazon tool that provides the intuitiveness. In the past, hiring was biased. The model learned the bias . Within the model, indicators of being female became indicators of a weaker application (Dastin, 2018). No one had to write a rule that said, “Prefer men.”

There is another reason why Bridges matters here. The Court of Appeal did not find, in fact, that the particular deployment had led to a racist match in the incidents before it. It found that the force had not taken reasonable steps to satisfy itself about bias despite a body of research showing higher error rates for women and for darker skinned people in a number of systems including the work later associated with Buolamwini and Gebru (2018) and NIST (Grother et al., 2019). The lack of investigation of a foreseeable risk was itself a breach of the public sector equality duty. I think that's the more important holding for policy. It makes you feel that you have a duty to look, not a duty to apologise after a bad game.

Two warning. First, disparate error rates are not the only equality concern. A system can be accurate and still reproduce an unjust baseline, which is closer to the Amazon story than the facial-recognition story. Second, the meaning of “fairness” is another issue introduced by criminal justice tools. Equal false-positive rates and equal calibration are not always achievable at the same time. I’m not going to pretend that a graduate paper settles that statistical conflict. I am going to argue that a sentencer who cannot articulate which definition of fairness was chosen has not met the reasons-giving part of equality before the law .

7. Freedom of Expression and Access to Information

Article 19 of the Universal Declaration and Article 19 of the ICCPR concern expression and the right to seek and receive information. Both halves. Both halves have platform moderation. As Special Rapporteur, David Kaye warned that automated moderation struggles with context—satire, counter-speech, local meaning—and that the burden of over-removal falls disproportionately on groups whose language is less represented in the training data (Kaye, 2018). This is in line with my reading of the platform literature. A classifier trained on majority examples will be less reliable on the speech of people at the margin (Douek, 2021; Sander, 2020).

Amplification is the other failure. Engagement Ranking will favour content that captures attention. Sometimes that material is false or hostile or both. The Myanmar investigations are the case that makes this an abstract media-studies point (UN Independent International Fact-Finding Mission on Myanmar, 2018). Facebook hate speech was one of the components of the arena in which the violence against the Rohingya happened. The report does not license a simple slogan of cause and effect. It does license a regulatory question: if a company develops a system that predictably rewards inflammatory content, is "we also remove some posts" an adequate account of its human rights impact?

I don't think the two errors offset each other. Over-removal and under-removal are both the product of the same architecture: probabilistic decisions on a scale that excludes individual judgement. Choosing the balance is a normative task. It cannot be solved by a better F-score alone. That's why transparency reports that measure takedowns, and say little about what the recommender promotes, feel incomplete to my mind. They measure one mistake and keep the other in the dark.

8. The Right to an Effective Remedy

Article 8 of the Universal Declaration and Article 2(3) of the ICCPR require a remedy that is capable of investigating, attributing and, where appropriate, redressing. This is where the other three rights either come into existence or remain on paper.

Example of jurisdiction is Clearview. A sanction in Paris, by itself, does not undo a database queried from elsewhere. Notice example is Amazon. Applicants harmed by the experimental tool would have had no way of knowing to complain. An internal review led to the company making it public and it became a Reuters story (Dastin, 2018). That's not a system of remedies." It's leaking." The third problem with platforms is that the developer, the deployer and a huge amount of content created by users are all in the causal chain. A claimant has to choose a respondent and then explain a model.

I have learned to see opacity as a covert rule of evidence. In practice, the time limit for bringing proceedings does not start until the person concerned can see that a model was used, as the claim is not foreseeable. Notice, therefore, is not a courtesy. That's part of the treatment. So is a collective mechanism; Algorithmic harm is often small per person, but large in aggregate . The recommendations also include representative actions, because a procedure tailored only to the individual claimant will under-enforce that pattern.

9. Comparative Legal Responses

The responses on the table now vary in strength and I have tried to show that difference instead of calling them all "governance."

At the international level, the Guiding Principles remain the benchmark for corporate responsibility, and B-Tech is the attempt to apply them to technology (UN Human Rights Council, 2011). UNESCO's 2021 Recommendation on the Ethics of Artificial Intelligence is broad – dignity, rights, environment – and not binding (UNESCO, 2021). The OHCHR's work on privacy in the digital era has put pressure on states on surveillance and biometric identification (UN Office of the High Commissioner for Human Rights, 2021). No one of these instruments, by itself, constitutes a court.

The text of the Council of Europe is different in form. On 5 September 2024, the Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, CETS No. 225 was opened for signature in Vilnius (Council of Europe, 2024). This is the first pact for this subject. It requires parties to ensure AI activities are compatible with rights, democracy and the rule of law, including risk and impact assessment at every stage of the AI system lifecycle. It will enter into force after ratification by five parties including at least three member states. That threshold is the legally relevant fact at the time of writing this paper. A treaty open for signature is not yet a treaty in force. Its obligations also leave a good deal of implementation choice to parties, and they bind states in their own use and in their regulation of the private sector, rather than binding companies directly.

The binding regional statute is EU Artificial Intelligence Act, Regulation (EU) 2024/1689, adopted on 13 June 2024 (European Parliament & Council, 2024). It's a risk based approach. Some practices are prohibited, including certain manipulative systems, certain biometric categorisation and, subject to narrow law enforcement exceptions, real-time remote biometric identification in public spaces. Many high-risk systems — including those used for employment, credit and law enforcement — have duties around data, documentation and oversight by humans. Article 27 mandates a fundamental rights impact assessment prior to first use of certain high-risk systems by bodies governed by public law, private persons providing public services, and deployers of certain employment and credit systems. I had expected from the public debate a general duty on every private developer. The text is more narrow. That narrowing is one reason my recommendations go beyond the Act rather than stop at it.

Domestic litigation still counts. Bridges shows Article 8 and equality duties can bite without an AI statute. The Clearview decisions illustrate how the cross-border provisions in the GDPR can be used. They also show the ceiling. The regulator can be right and not paid.

Table 1 provides a comparison of the instruments, as the prose comparison is easily muddled.

Table 1

Selected instruments and what they really bind

Instrument	Status	Bound To	What a student should not over-read
UN Guiding Principles (2011) and B-Tech	Soft law	States, as guidance; companies, as a responsibility	Not a cause of action by themselves
UNESCO Recommendation (2021)	Non-binding	Member states, encouraged	Language of ethics is not a cure-all
GDPR (2016), as it applies to biometrics and scraping	Binding in the EU/EEA, with extraterritorial hooks	Controllers and processors	Fine is not the same as globally deletion
EU AI Act, Regulation 2024/1689	Compulsory; staggered application	Providers and deployers, by risk level	Article 27 is not a general private-sector obligation

Council of Europe Framework Convention, CETS 225 (2024)	Opened for signature; not yet in force	States parties, once in force	Signature not constituting entry into force

Note. Author's summary of instruments used in this section. The text is not replaced by the cells.

This section includes one timing point, given the tendency of students to talk about the AI Act as if it has already been fully implemented. Regulation (EU) 2024/1689 adopted on 13 June 2024 was published in July 2024. The prohibitions and the duties on high-risk systems come into application on different dates, over a period of years. A paper written in 2026 can outline the duties, It can't consider them a settled enforcement record yet. I've tried to keep that distinction. The Council of Europe convention issues a similar warning, but from the other direction. It can be signed, and the European Union has authorised signature, without having the five ratifications needed to enter into force (Council of Europe, 2024). Sign language is a political fact. It is legal to come into force.

There is also a Global South gap that I should not leave implied. Most of the binding text I have discussed is European in origin. Facial recognition, welfare scoring and platform amplification are not European things. A framework that is serious in Brussels and paper-thin everywhere else will export costs of compliance and not export remedy. I have no room for a full comparative study of African, Asian and Latin American statutes. "The limitation is real. Recommendations that assume a data protection authority with GDPR-like powers will not be transferred to a state without such an authority.

10. Discussion

And across privacy, equality, and expression I see the same pattern. There is serious interference. A fair reading of existing law can already describe it as a rights problem. The institution that could have interrupted it comes later, and often only in one jurisdiction. Bridges is the hopeful case, because a court used old tools well. Clearview is the discouraging case, because some authorities used more recent tools, and the service was still available to debate jurisdiction. Myanmar lies in the middle: the damage was documented, the platform changed, and no one would claim the information environment was repaired.

The newer instruments focus on ex ante assessment . The AI Act, the Framework Convention and B-Tech all want someone to look before deployment, not just to sue after. I have that instinct too, I do. I am less comfortable with who sees. In the Act, the deployer evaluates. Much of the business-and-human-rights guidance is a company assessment. Self-assessment is a serious thing. It can also be a memorandum of the project. For biometric surveillance and for recommenders at the scale of a major platform, I don't think self-assessment is enough. That is a normative judgement. They do not, in themselves, prove that all internal review fails. What they do show, though, is that an internal review wasn't what brought Clearview, or Amazon's tool, or South Wales Police's equality failure, to public attention.

Let me be clear about what I am not arguing. I am not saying that each use of a model is a human rights violation, or that innovation and rights are opposed. A system that helps a radiologist find a tumour, or that translates a defendant's testimony, can serve rights. The cases of this paper are those where scale, opacity or a bad objective function cut the other way. If you're a graduate student who wants to make AI the villain, you're missing the legal question. It's about design and deployment and accountability, not whether software exists.

There is also a distributional point I don't want to lose. Those most likely to suffer bad matches, over-removal, and opaque hiring screens are often those least able to afford a test case. It has not bridged the remedy gap. A formula that works for a well-lawyered NGO in London or Paris, and not for an applicant who never knows a model was used. And it has moved it.

One last element is institutional capacity. Judgements can be carefully written, as in Bridges, but the next police force still has to read the judgement on its own. Regulators can fine a non-resident company the maximum the GDPR allows, as the CNIL did, and still be writing to a company that questions the premise of the letter. If impact assessment turns out to be the main instrument, it will fail in the same way if no one outside the deployer can ask for the assessment and compare it with the system that was actually switched on. I have begun to think of this as a problem of file-keeping. Human rights law is strong on standards. It's weaker on the boring machinery that makes a standard checkable: logs, watchlist criteria, training-data sheets and a person whose job it is to read them.

Graduate work on this subject can be honest about uncertainty there too, in the machinery. I can't tell from public record how often a live facial recognition alert is wrong in a given city. I can see an unasked force is already outside of Bridges. I don't know what portion of the 2016 vote was affected by micro-targeting. From the data, I see that it was collected without the consent required by the GDPR's predecessor regime and then the GDPR. The paper is based on that latter type of claim. Where the first sort would be necessary, I have mentioned it.

11. Suggestions

They are derived from the reading above. They are addressed to international organisations, states and developers. I have attempted to pair each one to a gap that the paper had actually identified.

States in a position to ratify the Council of Europe Framework Convention should do so, not wait until it has entered into force to legislate at home. Domestic law should mandate a human rights impact assessment before deployment of systems that pose a serious risk to privacy, equality or expression. The EU Act Article 27 is a start and a ceiling. It should be extended to private developers for the highest-risk uses — large-scale biometric identification and moderation or recommendation at platform scale — and the assessment should be open to independent check rather than left entirely to the deployer.

Second, enforcement cooperation needs to catch up to the Clearview file. Instead of parallel fines that remain unpaid, we should have coordinated orders and mutual recognition of findings. That would reduce the advantage of having no local assets. This isn't to say that it's easy. I think a stack of unenforced national penalties is evidence that the current arrangement is not working.

Third, the Bridges equality holding should be generalised. Public and private deployers of biometric identification, hiring, credit and similar high stakes systems should be required to test for disparate impact before deployment and revalidate afterward. The work is to look, and to record the look." “We did not know to ask” is a feeble reply given research already in the public domain (Buolamwini & Gebru, 2018; Grother et al., 2019).

Fourth, notice should be proactive. When an automated system makes or materially impacts a consequential decision in the employment, credit, or similar setting, that person should be notified, at a level of detail sufficient to contest it. Article 2(3) of the ICCPR is not a description of a right to be guessed at.

Fifth, large platforms should be subject to a human rights impact assessment and disclosure requirement for recommendation systems, including amplification of violent, discriminatory or seriously false content. Takedown totals don’t answer the Myanmar question. It does amplify.

Sixth, the Guiding Principles and B-Tech guidance for AI development should be made binding, either in a future instrument or through expansion of existing mandatory due diligence statutes in some European systems. A responsibility to respect which cannot be enforced will remain a speech in practice.

Seventh, the remedy must be usable. That means legal aid, easier complaints before data protection and equality bodies, and representative actions for harms that are small individually and widespread in aggregate. The technical file is part of the case and here specialised advice is important.

Eighth, surprisingly much of what we know came from journalists, whistleblowers and independent researchers not regulators. Rules should be narrowed to limit a terms-of-service clause or an anti-circumvention claim to shutting down good faith auditing, with protections against abuse. I mention this because a legal framework that depends on insiders to become visible, and then punishes the insider, will stay blind.

Ninth, border and welfare systems need the same notice and challenge rights as hiring tools and they need them more urgently because the person affected may be detained, refused entry or left without a payment while the challenge is prepared. A score used at a border should state in a language the person understands that a score was used, what the key factors were and where a review can be requested before removal. I don't have such a case already requiring this in those terms. I infer it from the remedy provisions already cited, applied to a setting the famous cases have under-described.

10. A publication rule for high-risk public deployments should be part of any duty of impact assessment. A summary of the purpose, the data, the groups for which unequal error was tested, and the human

supervision arrangement can be published without publishing the model. Sometimes it's justified to keep things secret. It is less often justified in the form of a blanket. Bridges was in part based on the public's inability to anticipate the rules. A secret policy cannot be predicted.

I would also add a note about what not to recommend. Chemical or purely technical fixes, a fairness constraint dropped into a loss function, are no substitute for the legal steps above. They can shrink a measured gap. They can't tell the claimant who is responsible. They can't make a forum. I read policy papers that end with a toolkit and I want the toolkit linked to a duty. Otherwise it's advice.

12. Conclusion

I started with a misfit between the actor human rights law anticipates and the system it increasingly faces. The cases don't show that privacy, equality, or expression have become irrelevant. They demonstrate that those rights can be interfered with by model, watchlist or ranking rule and that attribution and remedy lag. Bridges demonstrates that the old law can work. The Clearview fines are proof that even if a legal theory is correct, it cannot change the service. Amazon episode is proof that discrimination need not have a discriminating instruction. The platform cases are proof that speech governance is now partly a design choice about what gets seen.

The AI Act and the Council of Europe convention are the most serious attempts to date to put rights front and center in the governance of these systems. Both are incomplete in ways that the paper has tried to specify: phased application and a limited assessment duty in the Act; signature without entry into force, and state-mediated duties, in the convention. Self-assessment is still the default. Cross-border enforcement is the weak point. People are still mostly only aware of a model if someone else exposes it.

The position I can defend, then, is a modest one. Innovation doesn't have to be treated as an exception to rights, and rights don't have to be treated as a reason to stop every system. What must change is the assumption that a named decision-maker, a subsequent lawsuit, and a voluntary corporate policy will be enough. They were not numerous enough in the files this paper examined.

12. Conclusion

I started with a misfit between the actor human rights law anticipates and the system it increasingly faces. The cases don't show that privacy, equality, or expression have become irrelevant. They demonstrate that those rights can be interfered with by model, watchlist or ranking rule and that attribution and remedy lag. Bridges demonstrates that the old law can work. The Clearview fines are proof that even if a legal theory is correct, it cannot change the service. Amazon episode is proof that discrimination need not have a discriminating instruction. The platform cases are proof that speech governance is now partly a design choice about what gets seen.

The AI Act and the Council of Europe convention are the most serious attempts to date to put rights front and center in the governance of these systems. Both are incomplete in ways that the paper has tried to specify: phased application and a limited assessment duty in the Act; signature without entry into force, and state-mediated duties, in the convention. Self-assessment is still the default. Cross-border enforcement is the weak point. People are still mostly only aware of a model if someone else exposes it.

The position I can defend, then, is a modest one. Innovation doesn't have to be treated as an exception to rights, and rights don't have to be treated as a reason to stop every system. What must change is the assumption that a named decision-maker, a subsequent lawsuit, and a voluntary corporate policy will be enough. They were not numerous enough in the files this paper examined.

References

Access Now (2018). Human rights in the era of artificial intelligence. Get Access Now.

Amnesty International. (2019). Surveillance giants: the business model of Google and Facebook threatens human rights. Amnesty International.

Angwin, J., Larson, J., Mattu, S., Kirchner, L. . (2016, May 23). Machine bias. ProPublica https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Authority Personal Data. September 3, 2024). Dutch DPA slaps €30.5 million fine on Clearview AI https://www.autoriteitpersoonsgegevens.nl/en/current/dutch-dpa-fines-clearview-ai

Bridges v Chief Constable of South Wales Police [2020] EWCA Civ 1058.

Buolamwini, J. & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research, 81, 1-15.

Cadwalladr, C., Graham-Harrison, E. March 17, 2018. In a major breach, Cambridge Analytica harvested data from 50 million Facebook profiles. The Guardian. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-elections

CNIL. 20 October 2022. Facial recognition: Clearview AI fined 20 million euros by the CNIL. https://www.cnil.fr/en/facial-recognition-cnil-fines-clearview-ai-20-million-euros

Council of Europe. (2024). Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (CETS No. 225). https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=22.5

J. Dastin (2018, October 10). Amazon drops secret AI recruiting tool that showed bias against women Reuters https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G?rss=1&feed=rss

Douek, E. (2021). Regulating Online Speech: From Posts-as-Trumps to Proportionality and Probability. Columbia Law Review, 121, 759–834.

European Parliament and Council. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation) Official Journal of the European Union, L 119, pp. 1–88.

Council & European Parliament. (2024). Regulation (EU) 2024/1689 establishing harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union L 2024/1689

Warranty of Protection of Personal Data. (February 10, 2022) Italian SA fines Clearview AI €20 million for facial recognition. https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9751362

Grother, P., Ngan, M. & Hanaoka, K. . (2019). "Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects." NISTIR 8280. National Institute of Standards and Technology (NIST) https://doi.org/10.6028/NIST.IR.8280

Hellenic Data Protection Authority July 13, 2022. Ruling to fine Clearview AI 20 million euro.

Information Commissioner's Office, (23 May 2022). ICO fines facial recognition database company Clearview AI Inc more than £7.5m https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2022/05/ico-fines-facial-recognition-database-company-clearview-ai-inc/

Kaye, D. (2018). Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression (UN Doc. A/73/348). United Nations

Kaye, D. (2019). Speech police: The global struggle to control the internet. Columbia Global Reports.

Klonick, K. (2018). The new governors: people, rules and processes governing online speech. Harvard Law Review, 131(6), 1598-1670.

Nemitz P. (2018). Constitutional democracy and technology in the era of artificial intelligence. Philosophical Transactions of the Royal Society A, 376, 2133, 20180089. https://doi.org/10.1098/rsta.2018.0090

S and Marper v United Kingdom, App. 30562/04 and 30566/04 (Eur. Ct. H.R. Dec. 4, 2008).

Sander, B. (2020). Freedom of Expression in an Age of Online Platforms The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation Fordham International Law Journal, 43(4), 939–1006.

Suzor, N. . (2019). Lawless: The hidden laws shaping our digital world. Cambridge University Press.

United Nations Human Rights Council. (2011). Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework (HR/PUB/11/04) – United Nations.

UN Independent International Fact-Finding Mission on the situation in Myanmar. (2018). Report of the detailed findings of the Independent International Fact-Finding Mission on Myanmar (UN Doc. A/HRC/39/CRP.2). United Nations.

Office of the UN High Commissioner for Human Rights (2021). The right to privacy in the digital era: Report of the United Nations High Commissioner for Human Rights (UN Doc. A/HRC/48/31) United Nations

UNESCO. (2021). Recommendation on the ethics of artificial intelligence. UNESCO.

Wachter, S. (2020). Discrimination by association and affinity profiling in online behavioural advertising. Berkeley Technology Law Journal, 35(2), 367–430.

Yeung, K. , Howes, A. , Pogrebna, G. . (2020). AI governance through human rights-based design, deliberation and oversight. In: M. D. Dubber, F. Pasquale, & S. The Oxford handbook of ethics of AI (pp. 77–106). Eds Das, Oxford University Press.

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. Public Relations

// DECLARATIONS

▸ **Funding**	This research received no specific grant from any funding agency.
▸ **Conflict of Interest**	The authors declare that they have no known financial or personal conflicts of interest that could have influenced the work reported in this paper.
▸ **Data Availability Statement**	No new data were created or analysed in this study. The article is based on legislation, regulatory and policy instruments, reported judicial decisions and published scholarship cited in the references.
▸ **Ethical Statement**	Ethical approval was not required because the study is a doctrinal and comparative policy analysis of publicly available legal texts, judicial decisions, policy documents and published literature and did not involve human participants.
▸ **AI Use Disclosure**	Generative artificial intelligence tools were used in a limited and assistive capacity during the preparation of this manuscript, primarily for idea development, language refinement, and grammatical correction. The authors retain full responsibility for the originality, accuracy, analysis, interpretation, and final content of the manuscript.

// JOURNAL POLICIES *JuriSocial Journal of Criminology & Criminal Justice*

01 OPEN ACCESS & LICENCE
All articles are freely available on publication under CC BY-NC 4.0. Readers may share and adapt the work for non-commercial purposes with proper attribution.

02 COPYRIGHT
Authors retain copyright and grant JJCCJ the right of first publication.

03 PEER REVIEW
Every research article undergoes double-blind peer review by at least two independent reviewers.

04 PUBLICATION ETHICS
JJCCJ follows the Committee on Publication Ethics (COPE) Core Practices and guidelines.

05 PLAGIARISM POLICY
All submissions are screened with similarity-detection software; plagiarism in any form leads to rejection.

06 GENERATIVE AI POLICY
AI tools cannot be listed as authors. Any substantive use must be disclosed; authors remain fully responsible for the content.

07 CORRECTIONS & RETRACTIONS
Errors are corrected through published corrigenda; retractions follow COPE guidelines.

08 ARCHIVING & INDEXING
Content is preserved through [PKP Preservation Network / LOCKSS]. DOIs are registered with Crossref.

09 PUBLICATION CHARGES
No article publishing fee charged.

10 DISCLAIMER
The views expressed are those of the authors and do not necessarily reflect the views of the editors or the publisher.

SUBMISSIONS & ENQUIRIES [editor@journal-domain.org] ◆ [https://journal-website-url/]