https://doi.org/10.65025/ICAIC26066d

From Voluntary AI Ethics to Mandatory Compliance: Addressing Enforcement Gaps in Human Rights and Security Governance

Xena Dominique NGOYA DAVİD
Final International University, Cyprus.
xena.tsimi@final.edu.tr
https://orcid.org/0009-0009-5616-3072

Abstract

In 2020, for instance, Robert Williams was illegally arrested after an AI face recognition attempt failed, one of the many examples that shows that AI has far outstripped efforts to control it. Rising set of ethical rules based on AI responsibilities, developed by governments, international organisations and business actors, while generally voluntary and binding, remain largely ignored in the context of situations where AI systems directly affect human rights and security. There's nothing wrong with principles, it's the mechanism to enforce them. There are few actual rules, if any. Promises are made, but then broken. Ethical guidelines are still in the minority in terms of techniques for governance of AI, as Facebook found out in its involvement with the Cambridge Analytica data fiasco, where the enforced teeth of the General Data Protection Regulation were what proved most clattering. This lesson has not been learnt sufficiently. There is no risk-free mandatory compliance. However, the voluntary approach has failed to work out as intended and these risks need to be weighed against the harm that can be caused by AI if it is not governed. In the opinion of this paper, AI governance should not remain aspirational, it must be enforceable by law. It calls for a policy framework based on mandatory due diligence obligations, transparency, independent oversight and effective sanctions that "actually bite". It also argues that coordinated global enforcement models like the EU AI Act, are required.

Keywords: AI governance; human rights; Enforcement Gaps; Security governance; mandatory compliance; voluntary ethics; EU AI Act; due diligence; binding regulation; accountability.

1. Introduction

Artificial intelligence (AI) technologies are being adopted by public and corporate sectors way faster than the mechanisms designed to control them. Artificial intelligence systems are currently involved in decisions with major repercussions for individuals and communities, including healthcare, migration control, financial services, employment, and criminal justice (Russell & Norvig, 2021).Many ethical frameworks try to lead responsible AI . However, a survey of 84 AI ethics guidelines has found that while many of them mention comparable values, they are significantly different in practice , and most do not have meaningful enforcement or punishment systems (Jobin et al., 2019) . The gap between ethical promises and actual behaviors is becoming increasingly visible in contexts where AI systems have a direct impact on human rights and security, and more costly to those who bear the consequences.

A popular example is the 2020 wrongful arrest of Robert Williams, in which law enforcement officers, using a faulty facial recognition system, jailed an innocent individual in front of his family (Hill, 2020). Williams was held for thirty hours before the mistake was discovered. There was no supervision method to catch it, and no voluntary moral framework to prevent the error. What is clear is that none of this was unforeseen, with academics identifying accurate differences in facial recognition algorithms years before the Williams's case (Buolamwini & Gebru, 2018). UNIDIR has also warned that in security contexts, the opacity of many AI

systems makes it impossible to verify their trustworthiness, detect mistakes or establish accountability when harm happens (Grand-Clément, 2023). The Williams case did not expose a gap unknown to anyone, but a gap no one had been obliged to close.

Again the GDPR records tell a different story. The harvesting of personal data without informed consent of tens of millions of users across several jurisdictions in the Facebook-Cambridge Analytica issue has shown that no ethical norm has ensured responsibility for such activities (Cadwalladr & Graham-Harrison, 2018). But Meta Platforms was slapped with a €1.2 billion fine by Ireland's Data Protection Commission on May 22, 2023. Enforced legal systems have shown to be able to provide accountability where voluntary agreements fail, as the highest GDPR fine to date was against Ireland for illegal transfers to the USA (Irish Data Protection Commission, 2023; European Data Protection Board, 2023). So, why has the lessons learnt for accountability in data protection not been learnt when it comes to AI? Not yet, though. There is a general lack of hard-law measures to govern AI. The OECD Recommendation on Artificial Intelligence is the first international standard on AI and outlines five values-driven principles for the responsible use of trustworthy AI, but are not binding (OECD, 2019). The Recommendation on the Ethics of Artificial Intelligence from UNESCO, adopted by acclamation by all 193 Member States in November 2021, is equally devoid of legislative measures (UNESCO, 2021). UNIDIR's research on AI governance in security and military contexts has found that the multilateral AI discourse is largely centered on voluntary frameworks, resulting in significant accountability gaps for states deploying opaque AI systems (He & Anand, 2023; Puscas, 2023). The need to regulate AI have not only been felt by policymakers, academics and intergovernmental organizations. On 25th May 2026, Pope Leo XIV's first encyclical, Magnifica Humanitas (Magnificent Humanitas), sent to Catholics and "every person of goodwill" states that AI should be based on respect for human dignity and not focus power in the hands of a few stackholder and it deems self-operating weapons to be morally incompatible. The Magnifica Humanitas also demands explicit binding international regulations like those regulating nuclear arms and It defines governance to be: traceability of decisions, meaningful human control over lethal action and enforceable international rules. (Leo XIV, 2026). Voluntary approaches to AI Regulation have been attempted. The paper contend that ethical rules are not the issue at hand when it comes to AI governance: it is the lack of effective enforcement of them. There is an increasing need for compulsory compliance systems on the basis of human rights obligations and security concerns. However, these concerns should be balanced against the actual damage that can be caused by unchecked systems and these are no longer theories. Independent monitoring, real investigative power and imposed sanctions are needed, not as cumbersome administration but as an essential component to real government (European Parliament, 2024).

Domestic enforcement is not enough to accomplish this. The emergence of AI systems has been fueled by data from across the globe, implemented across borders and potentially have harmful impacts that can only be mitigated by the combined efforts of multiple governments. When regulation is fragmented, monopolies result. UNIDIR has shown that there have been, in the past, only sporadic attempts to discuss AI in a security context in a multilateral forum, which center around one or two specific international mechanisms, thereby highlighting the need for more extensive and organised international engagement. This is a very significant difference. One of the examples of coordinated enforcement can be seen in the EU AI Act (He & Anand, 2023). It is not possible to build on it across jurisdictions as this will create a distinction between governance that is and governance that is effective. This paper takes up a basic issue in the governance of AI, namely the lack of translating moral pledges into obligations. It is based on the example of the unjust arrest of Robert Williams and the fact that voluntary frameworks do

not lead to harm, while legal regimes do. It also provides a compliance framework based on due diligence, openness, independent oversight, and effective sanctions. It argues most crucially, that the transnationality of AI applications in the field of security requires a worldwide law for AI applications. The soft law has been fragmented and has failed. Now there's a need for a coordinated and actionable international framework.

2. Literature Review

2.1 The Voluntary Ethics Paradigm: Widlyspread but Weak

AI ethics theories has experienced an incredible surge in the past ten years. There are hundreds of guidelines, principles and suggestions from governments, multilateral organizations, technology corporations and civil society groups on how to develop and implement responsible AI. There are some general principles that Jobin et al. identified in their analysis of 84 of such publications, which seems to be a global consensus on the nature of ethical AI: transparency, justice and fairness, non-maleficence, responsibility and privacy (Jobin et al. 2019). Nevertheless, this hasn't translated into compliance. While AI Ethics is a focus area with much promise, as Mittelstadt noted in response to the general optimism on this topic, Principles, by themselves, are not sufficient to ensure an ethical AI and without institutional mechanisms to implement the principles, the frameworks act as repurposing tools rather than governance tools (Mittelstadt, 2019).

This critique has gained tremendous traction in the academy. Calo provided a general map of the critical policy dilemmas that AI raises, which pointed to the governance problems generated by the lack of binding regulatory frameworks. Since then, critics have contended that this environment allows the sector to use ethical discourse while continuing with damaging actions. (Calo, 2017). Similarly, Dafoe's seminal research agenda on governance cautioned that voluntary frameworks are inherently plagued by credibility problems: they depend on the good will of players whose commercial motives sometimes run opposite to the principles that these frameworks promote (Dafoe, 2018). What is notable about this collection of work is not its originality but its consistency: for almost a decade, researchers have pointed to the same fundamental defect, and policymakers have mostly refused to act upon it.

The gap addressed by this research is therefore not a gap in principle articulation where work has been done. It's a weak point in the enforcement structure. Great care has been taken to diagnose the problem in scholarship. What is missing is a clear, legally sound concept for what necessary compliance in AI governance should look like, how it should be structured across jurisdictions and what role international institutions like UNIDIR should play in its development and oversight.

2.2 Human Rights Frameworks and AI: A Promise Without Teeth

International human rights law provides a natural normative underpinning for AI governance. The rights that are directly impacted by the deployment of AI, namely the right to privacy, the right to a fair trial, the right to non-discrimination and the right to an effective remedy, are laid down in the Universal Declaration of Human Rights, the International Covenant on Civil and Political Rights and regional instruments such as the European Convention on Human Rights (United Nations, 1948; United Nations, 1966). The problem is that these instruments were designed for a pre-digital context, and have been applied to AI only by interpretive extension, not by legislative compulsion.

In several resolutions, the UN Human Rights Council has identified the human rights implications of AI, including that it can be used to reinforce and amplify discrimination, to

infringe on privacy and to violate due process (UN Human Rights Council, 2021). However, these resolutions are not binding but a declaration. As Raso et al. have pointed out, the use of human rights law in the context of AI is thus far disjointed and sporadic, as states apply various human rights concepts in a selective manner without a consistent approach (Raso et al., 2018). The Williams case fits into this category: the use of a racially biased facial recognition technology by police in a nation that supposedly upholds the rule of law that does not have any kind of clear and binding guidelines to curb its misuse (Hill, 2020; Buolamwini & Gebru, 2018). The tension was recognized by the Ethics Guidelines for Trustworthy AI of the High-Level Expert Group on AI who recommended the implementation of basic rights impact assessments for high-risk AI systems (High-Level Expert Group on AI, 2019). However, the guidelines were not mandatory. This article then closely examines the EU AI Act's action steps to put some of these concepts into legally binding action. Its geographic scope and architecture of enforcement create large gaps, particularly in the context of security where the primary actors are not private companies but state actors, who are also the primary agents to deploy risk-bearing AI systems.

2.3 What does the GDPR teachs about Data Protection governance model?

The most helpful framework for AI to be held accountable is the GDPR. The GDPR, which came into effect in 2016 and was first tested in numerous high-profile enforcement actions, has demonstrated that binding legal rules (and independent enforcement and substantial financial sanctions) are effective in establishing accountability in the governance of technologies where voluntary measures have proven ineffective (Voigt & Von dem Bussche, 2017). The fine, the largest under the GDPR since 16 July 2020, was imposed on Meta Platforms Ireland Limited for transferring personal data to the US, under SCCs. The Irish Data Protection Authority (IDPA) decided this after the EDPB issued a binding dispute resolution ruling on 13 April 2023 (European Data Protection Board, 2023). One of the Brussels Effect's strengths is the analysis of how the GDPR's rules, designed for the European market, can become global standards. As far as the has, it has been able to spread European data protection standards globally as corporations are inclined to accept standard compliance policies, as opposed to those varying by jurisdiction (Bradford, 2020). This pattern has global compliance impacts that have very important consequences for the governance of AI: strong compliance effects can stem from a regional compliance framework, if it starts locally. In a similar way, the EU AI Act is currently shaping the evolution of AI development practices by transnational companies (European Parliament, 2024). The GDPR approach to AI governance does, however, not easily translate. Data protection law works with very clearly defined categories personal data, processing, permission which are amenable to legal specification. AI governance has to deal with a much more complicated and quickly changing technological landscape, where the dangers associated with any particular system rely largely on context, deployment environment, and interaction effects that are hard to predict during the legislative writing stage. The GDPR approach to the governance of AI is not easily translatable, however; data protection law is based on very well-defined notions like personal data, processing, permission that are susceptible to legal specification. AI governance needs to deal with a far more complex and evolving technological environment, and the risks of each specific system are highly context-dependent, tied to its environment, and dynamic, meaning the context in which any AI interacts can have an impact on its risks. As argued by Wachter et al. the GDPR is “toothless” as it “lacks precision in its language and does not explicitly define rights and safeguards from automated decision-making, indicating the lack of a “right to explanation” under GDPR applied to AI. (Wachter et al., 2017). Therefore, specific AI regulation should be more flexible and principle-based than data protection law, while retaining the enforcement backbone that makes the GDPR effective. This essay supports and elaborates this approach.

2.4 The EU AI Act: A Binding Framework and Its Limitations

Officially came into force on 1 August 2024 with penalty of up to €35 million or 7% of worldwide turnover for the most serious breaches, the EU AI Act is the most important piece of AI governance legislation to date. Its risk-based approach, which classifies AI systems into unacceptable, high, limited and low risk categories and imposes duties commensurate to the amount of risk, presents a practical model for obligatory compliance that other jurisdictions can adapt (European Parliament, 2024). The Act calls for conformity evaluations, transparency requirements, human oversight and registration in a public database for high-risk systems, including in law enforcement, border control and essential infrastructure. Breaches of high-risk AI can result in fines of up to €15 million or 3% of turnover and violations related to prohibited AI systems can result in a fines of up to €35 million or 7% of global annual turnover, which is greater. This is a good step forward. However, it is limited and the current research is beginning to highlight this. First, its reach, though vast, doesn't cover systems deployed by third country governments for national security purposes, where the harm of AI is most extreme and least public(Veale & Borgesius, 2021). Secondly enforcement depends on national market monitoring authorities whose capacity, independence and resources vary greatly between member states. Third, and most important for the case being made here, it is operating inside a single geographical framework. Any AI system used across borders in military operations, international surveillance, or transnational security cooperation is outside the reach of binding oversight by any one authority.

UNIDIR's research on AI in security contexts has consistently documented this gap, repeatedly pointing to the absence of multilateral frameworks that can govern the use of AI in defence and security operations at the same level of accountability that civilian applications are beginning to attract (He & Anand, 2023; Puscas, 2023). This study contends that the bridging of this gap necessitates a move from regional frameworks to a globally binding legal instrument. This argument is elaborated in the analysis sections that follow.

2.5 The paper contribution

Three important, but insufficiently integrated bodies of work have been produced by the extant literature: study on the limitations of voluntary AI ethics frameworks; scholarship on the application of human rights law to AI; and scholarship on the EU AI Act as an emergent compliance model. What is absent is a synthesis, a clear argument for how these elements might be brought together into a proposal for mandatory AI compliance based on human rights duties, operationally specific and applicable worldwide, even in security scenarios. That synthesis is provided in this work. It goes beyond simply claiming that voluntary ethics has failed but it argues for what must replace it, and why the construction of that replacement matters as much as its existence.

3. Research Methodology

In this Article, a doctrinal and comparative legal methodology, with policy analysis, will be applied. The doctrinal approach involves a detailed study of primary sources including the EU AI Act, the GDPR, international human rights instruments and UNIDIR documentation to understand the existing legal landscape in which AI governance is currently conducted and what gaps the mandated compliance mechanisms need to fill. It compares it with developments in various jurisdictions in its comparative dimension, with a special focus on the contrast between the binding regulatory approach of the EU and the largely soft-law frameworks found in other jurisdictions, including the US and the international arena. The choice of doctrinal method is intentional as AI governance is primarily a legal challenge, concerning the distribution of rights,

duties and accountability, and legal analysis is thus the correct main tool. At the same time, pure doctrinal study risks abstraction from real-world contexts in which AI systems create harm. In this way the study contributes to legal analysis, policy analysis drawing on recorded incidents of AI-related harm and the institutional literature on governance efficacy in order to anchor normative ideas in actual reality.

The key drawback of this study is that it is concerned with the supply side of governance, namely the creation of legal frameworks, rather than their implementation and enforcement in practice. Empirical analysis of how binding AI compliance regimes actually function in the countries who have accepted them is still a key field for future research, as is systematic investigation into the political economy of international discussions on AI governance. We are upfront about these limitations. This study contributes an architecture rather than actual work, and should be regarded as a platform for reform rather than an evaluation of existing outcomes.

4. Analysis

4.1 Why Voluntary Frameworks Have Structurally Failed

The failure of voluntary AI ethics frameworks is not incidental, it is systemic. Three properties of voluntary frameworks make it hard to control high-risk environments effectively using them. First They do not constitute a legal obligation. There is no binding obligation for an organisation to follow an AI ethical framework, there is no liability for the organisation if they fail to abide by it, nor is there a penalty for breaking it. Secondly, there is no independent verification procedure. No body has the authority to review adherence to a framework's promises, there's no right of audit, and there's no reporting obligation to the public. Third, they give no recourse for those injured by non-compliance. Sometimes the ethical framework of an AI system is only voluntary, which means that if rights are violated by an AI system, it does not give a person a basis on which to act. The three failures are all crystallised in the Williams case. Detroit Police Department did not have a legal obligation to use facial recognition technology. It had no independent organisation to assure its deployment met any minimal standards of accuracy or control. When Williams was misarrested, there was no early legal avenue provided by AI governance law that led to accountability; only civil rights constitutional litigation. (Hill, 2020; ACLU,2021). The voluntary ethics guidelines which were supposed to be followed in law enforcement use of AI were not relevant to his situation at that moment. This is a structural study which directly concerns reform. Making voluntary mechanisms stronger will not suffice if the mechanisms are not precise, if they are not widely accepted, or if they are not really endorsed. It isn't the content of voluntary structures, it's their legal nature. The duty needs to become more tangible: it must be a duty, not a choice; a duty that needs to be assessed rather than self-assessed; a duty that is sanctioned rather than a duty without consequences.

4.2 The rationale behind mandatory due diligence obligations

The first and fundamental component in any mandatory compliance framework for AI is legally binding obligations for developers and deployers of high-risk AI systems to exercise due diligence. Many scholars and experts have discussed the scope and nature of due diligence obligations in international law, as well as in other areas such as environmental law, human rights law and corporate law, and this concept has been incorporated in the context of AI (Ruggie, 2011). An obligation to conduct due diligence would involve organisations engaging in the development or deployment of high-risk AI systems in identifying, assessing and preventing, and addressing foreseeable risks to human rights and security, before and after the deployment, and of monitoring and addressing harms if they occurred during operation.

This is a partial illustration of the conformity assessment requirements of the EU AI Act for high-risk systems, which call for systematic risk assessments and the development of risk

management systems prior to placing a system on the market (European Parliament, 2024). They have three restrictions however. The conformity assessment process is largely self-assessed with the developers conducting their own assessments against harmonised standards and only third party assessment where there are limited circumstances. The definition of "high-risk" is very inclusive, but excludes a lot of applications of AI by state actors in the security spaces. Moreover, the territorial scope of the Act does not include the use of AI by non-EU organizations in non-EU contexts, even if this has an impact on EU persons or is connected with EU institutions. All three restrictions would have to be overcome first by developing a standard that is applicable worldwide. It would require an independent third-party evaluation for high-risk systems, would have a very wide definition of high risk, would include state security applications and would have extraterritorial jurisdiction, through mechanisms comparable to those established in international anti-corruption laws and laws on supply chain due diligence. Examples of trans-border laws with binding legal due diligence obligations include the French Duty of Vigilance Law (2017) and the German Supply Chain Due Diligence Act (2021) which include civil liability for non-compliance. This design is legal and just to transpose to the governance of AI.

4.3 High-Risk System Transparency Requirements

The second important aspect of a mandated compliance regime is transparency. There can be no independent monitoring without public accountability. "If you can't audit an AI system, you can't regulate it. The level of openness of AI governance is much deeper than meets the eye, however, as it happens at three levels: algorithmic openness, deployment transparency, and consequence transparency, all with different legal instruments. Algorithmic transparency is concerned with how all of these AI systems operate: the data that they were trained with, the logic they use to generate outputs, and the accuracy and error rates by different demographic groups. For high-risk systems the EU AI Act requires technical documentation and recording that enables post-hoc analysis of system behaviour (European Parliament, 2024). The practical importance of this problem was illustrated by Buolamwini and Gebru's Gender Shades research, which found that commercial facial recognition systems had error rates as high as 34.7% for darker-skinned females. This difference would have gone undetected had researchers not targeted their testing (Buolamwini & Gebru, 2018). Mandatory algorithmic transparency rules would make such discrepancies known before implementation, not post-harm.

Deployment transparency addresses the conditions under which AI systems are deployed: who deploys them, for what goals, against whom and under what authority. UNIDIR has identified the opacity of deploying AI in security contexts as a governance concern in particular. States regularly deploy AI-enabled systems in military and law enforcement contexts but do not reveal their capabilities, limitations or decision-making authority to the public (Grand-Clément, 2023). A mandatory transparency regime would require that high-risk AI deployments, including those carried out by state actors, be publicly registered in a format that is accessible to affected individuals and oversight bodies.

Outcome transparency is about the real-world impacts of deploying AI systems: what judgments are taken, who is affected, what mistakes are made. Mandatory outcome reporting similar to the adverse event reporting requirements in the regulation of pharmaceuticals would develop a systematic evidence base for determining whether AI systems are working as intended, and would enable early detection of systemic harms before they reach the scale of the Williams case.

4.4 Institutional setup of accountability: Independent oversight

Transparency rules only work if someone has the capacity to act on the information they generate. Thus, the third part of a mandated compliance structure is independent oversight bodies with the ability, resources and experience to oversee the deployment of AI, investigate complaints and issue remedies. The Irish Data Protection Commission's enforcement of the GDPR against Meta shows what independent oversight may deliver: a long, technical inquiry that led to the biggest ever data protection fine (Irish Data Protection Commission, 2023, May 22). The lesson is not simply that fines discourage. It is that reputable, independent agencies with real investigative powers can call even the most powerful technology operators to account. The major AI compliance oversight structure is the creation of national competent bodies in each member state, which are managed by a European AI Office (European Parliament, 2024). This is a reasonable paradigm for domestic AI governance, but has two shortcomings pertinent to the argument presented. Enforcement in the single market is uneven, with national authorities differing in capacity and independence. And the architecture is essentially domestic provides no mechanism for supervision of AI systems working outside national borders or deployed by foreign state actors.

A global architecture for AI governance would include entities that reach beyond national borders, similar to the International Atomic Energy Agency in nuclear control, or the Financial Stability Board in financial regulation. UNIDIR is ideally positioned to contribute to this architecture in the security sector. Its existing mandate, technical knowledge and contacts with member states provide a basis for enhanced supervision functions (He & Anand, 2023; Puscas, 2023). What is absent is the political resolve to expand UNIDIR's mission beyond research and dialogue to binding oversight, and the legal framework that would allow for such an expansion.

4.5 Enforceable Penalties: Making Sure Non-Compliance Costs You

The fourth element of a required compliance system is effective deterrence through sanctions. (2) This point is not as trivial as it may seem. Sanctions exist in many regulatory regimes, but are ineffective at deterring because they are too small relative to the benefits of non-compliance, too rarely imposed to constitute a credible threat, or structured in such a way that organisations treat them as a cost of doing business rather than a real constraint on behaviour.

This failure mode was deliberately avoided by the sanction structure of the GDPR, which fines up to €20 million, or 4% of global annual sales, whichever is higher (General Data Protection Regulation, 2016, art. 83), and ensures that penalties are scaled to the economic ability of the perpetrator. The EU AI Act takes a step further, with the inclusion of sanctions of up to €35 million or 7% of the global annual turnover for the most grievous breaches (European Parliament, 2024). Both are real improvements on the nominal fines that typified prior technology regulation.

To be effective, the penalty structure of a global AI compliance framework needs to meet three criteria. Sanctions must be commensurate with the injury done and the "economic ability to be sanctioned". They have to be enforceably applied consistently such that there are real expected penalties for not following them. They need to be accompanied by non-financial penalties such as mandatory remedial steps, disclosure of offences, and in very egregious cases, the banning of the deployment of AI systems. The financial and operational implications are what give the real bite to regulatory action, as was seen in the Meta case, where the fine was a hefty €1.2 billion, but the threat to cease transferring data to the United States was arguably the more significant (European Data Protection Board, 2023).

4.6 The Case for a Legally Binding Global Instrument

The four elements outlined above, relating to due diligence obligations, transparency standards, independent monitoring and enforceable sanctions can be implemented at the national or regional level. The EU AI Act is beginning to achieve just that. However well-designed, the domestic and regional implementation can not fill the security-implementation gap, because AI systems are either used by State actors internationally or, often, in territories that are not within the jurisdiction of any single State. The global nature of AI use in security settings results in the conditions of regulatory arbitrage, as Bradford puts it: actors who want to avoid accountability can do so either by placing operations in jurisdictions that are less well-governed, or by using AI systems through international partnerships that obscure accountability (Bradford, 2020). To address this, there is a need for an international instrument (a treaty or convention) that sets minimum standards for all states, adds binding obligations with enforceable mechanisms, and offers remedies for individuals and communities affected by non-conforming deployments of AI. It is not surprising that this is a suggestion. International law has evolved into binding instruments on weapons technologies the Chemical Weapons Convention, the Ottawa Treaty on anti-personnel mines and the Arms Trade Treaty which are all defining global standards, developing a process of verification and setting consequences in case of non-compliance (Chemical Weapons Convention 1993, Ottawa Treaty 1997, Arms Trade Treaty 2013). The governance problem provided by AI in security situations is similar: a powerful technology with great potential for harm, adopted by state and non-state actors across borders, in contexts in which domestic control is inadequate. The architecture of arms control law as a suitable blueprint for AI governance has not been fully engaged with in existing study.

UNIDIR's research has started to map out this picture, charting the current state of play of AI governance in the military domain and the international involvement required to solve it (He & Anand, 2023; Grand-Clément, 2023; Puscas, 2023). This study aims to achieve what the research has yet to do, namely, to establish the normative and legal case for why a binding instrument is not simply one choice among many, but the necessary end point of any meaningful AI governance reform program.

5. Critical Evaluation

To be honest, the approach described in this paper is a bit ambitious. It's not easy to get governments to agree on mandated AI standards, and the challenges are serious enough to warrant honest study, not to be set aside.

The most obvious problem is political. States really do have to sign up to bind international agreements and the countries with the most powerful AI projects, the United States, China and Russia, have made it quite apparent they do not want anyone teaching them what to do with their AI systems. It's not a guess, as in October 2025, the US and Russia vetoed a UN resolution on responsible military AI a resolution the US itself had put forward two years earlier on the grounds that global surveillance would jeopardise national sovereignty (Afina, 2026). China has been more quiet, attending international meetings and accepting declarations but also promoting its own independent governance agenda centered upon its interests (Caroli & Mande, 2025). However real political resistance is not the whole story. As mentioned in the introduction, Pope Leo XIV's Magnifica Humanitas is a quite fascinating example of the evolution of the need for binding governance of AI (Leo XIV, 2026). It is important to note not only who said it here but what it does to the political calculation. The US and Russia can argue, with some plausibility, that their objections to a UN resolution on military AI are related to sovereignty or a technical issue or a timing issue. It becomes more difficult to maintain that

notions of "voluntary ethics" have been shown to be ineffective and rules to be enforced, when the same sentiment is being echoed independently by international security researchers, regional legislators, and the head of the world's largest religious institution. Yet it remains to be seen whether Magnifica Humanitas will continue to impact the governance of AI in the long term. But what, on the other hand, cannot be argued is that the constituency for binding regulation is now too large and too diverse to be called idealistic.

There is actual and systemic resistance and there is the issue of technology. Governments are unable to catch up with the speed of AI. While the European Commission did not mention general-purpose AI models in its proposed AI Act of 2021, the general-purpose AI was the main concern in the entire discussion within the last eighteen months (Veale and Borgesius, 2021). Any attempt to control some technology by name will be obsolete even before it comes into existence. Also the economic issue: compliance is costly and those costs are likely to be the greatest for smaller organisations and less developed countries, prompting the unpleasant thought that strict rules internationally could reinforce the power of the biggest organisations and richest countries. (Vinals Musquera & Babwah Brennen, 2026). It is important to deal with these issues and not take them for granted. On the political side: As is well known, great powers walk away from international agreements and it has never been the end of the story. The United States resisted until a late stage in the process in getting the right to retaliate with chemical weapons, but it did (OPCW, 2024; Arms Control Association, 2024). It took the Chemical Weapons Convention 13 years to negotiate. The Montreal Protocol on ozone depletion had to face stiff opposition from chemical sector lobby groups and a handful of sceptical governments, but was the most successful binding environmental treaty ever negotiated, boasting 197 countries that have signed up to it, and a compliance rate of 98% (UNEP, 2024; Rapid Transition Alliance, 2019). These pledges were made not because they were the product of good intentions, but because of increasing evidence of damage that made it “fatal” to leave them to do nothing. It’s the same with AI. There is increasing awareness of potential threats to safety posed by AI in the military and security domains. There is increasing evidence of potential threats of AI safety in military and security sectors (Puscas, 2023). Wrongful arrests due to racial bias in facial recognition technologies (Hill, 2020; Buolamwini & Gebru, 2018), biased choices in hiring, credit and criminal sentencing (Raso et al., 2018). And this is the kind of proof that alters political calculations over time.

It is essential defining the Paris Agreement here, as it is often cited as a success story of international cooperation on a complex subject. In a way, it is. The event was held, however, because the United States, under President Obama, had pushed for a framework that had no legally binding limitations on emissions in order to avoid having to seek approval from Congress (Doniger, 2015). The result is an agreement that countries may and do walk away from with no legal consequence. This is not the model of this paper. The CWC and Montreal Protocol are the right analogies and they are the right analogues because they established real rules with real consequences for non-compliance.

On the technical challenge, the answer is not to avoid regulation, but to regulate differently. Rules about what a technology must do tend to get out of date fast. Whatever the underlying technology may change, regulations indicating what effects a technology must not cause harm to humans, lack of transparency, absence of human control remain applicable. This is the approach recommended here and it is the same logic that drives the risk-based approach of the EU AI Act (Mittelstadt, 2019; Veale and Borgesius, 2021; European Parliament, 2024). Once again, the economic challenge is illuminated by the Montreal Protocol. It established a Multilateral Fund to help developing countries satisfy their compliance obligations and phased

deadlines to provide poorer countries more time to adapt (UNEP, 2024). There's no reason a global AI framework couldn't do the same.

What I don't know is how we get from here to there. It is difficult to adopt a binding global instrument, as international political processes are slow, easy to stop and can be easily bent by the very persons that they are meant to govern (Dafoe, 2018; Calo, 2017). Given the EU AI Act, a more realistic short-term alternative is likely to be to give UNIDIR a stronger oversight role and push for regional frameworks that can gradually converge, which may well achieve more in the short term than an ambitious global effort that goes nowhere (He & Anand, 2023; Puscas, 2023). The study of the Brussels Effect by Bradford demonstrates that a well-designed regional regime can have an effect on global behaviour purely by market pressure, without being formally in control of other jurisdictions (Bradford, 2020). It's the what that matters. Here it needs to play its role in the big picture. But getting there slowly isn't a failure it's probably simply the way things goes.

6. Conclusion

We began this article by making an observation that is essential, but often overlooked - the ethical commitments and actual practice in AI governance is not a matter of principles but one of enforcement. It's not that no one thought facial recognition technology should be accurate and fair when it comes to misidentifying the wrong person. No one thought facial recognition technology shouldn't get it wrong, misidentify the wrong person when it comes to that. He was falsely arrested because the Detroit police department had no obligation to ensure that the system it used met any standards, and no one could prevent or repair the damage that was inflicted to him. In this paper have attempted to bridge the gap between articulation and accountability. The analysis has produced three main results. Structurally, the voluntary AI ethical frameworks are unable to guarantee compliance, they are not poorly designed, just that they do not impose any legal obligations, no independent verification and no redress, in high-risk scenarios. The second, with the GDPR experience is that binding legal frameworks can provide real accountability in technology governance and can serve as a model for AI governance to learn from and build on. Third, the transnational dimension of the use of AI in the security domain also makes the absence of effective minimum standards, independent oversight and enforceable sanctions in domestic and regional legal frameworks an obstacle to a serious reform agenda. There's not a short time interval between here (actual situation) and there (the goal). But the way is clear. There is a growing body of systematic evidence of voluntary ethics failure. It is not idealism, it's law that is needed for AI governance.

References

Primary Sources

1. Arms Trade Treaty, opened for signature April 2, 2013, 3013 U.N.T.S. 269. https://www.un.org/disarmament/convarms/att/

2. Chemical Weapons Convention, opened for signature January 13, 1993, 1974 U.N.T.S. 45. https://www.opcw.org/chemical-weapons-convention

3. Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on their Destruction (Ottawa Treaty), September 18, 1997, 2056 U.N.T.S. 211. https://www.apminebanconvention.org

4. European Parliament. (2024, March 13). *Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).*

5. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

6. Gesetz über die unternehmerischen Sorgfaltspflichten in Lieferketten (Lieferkettensorgfaltspflichtengesetz — LkSG) [Supply Chain Due Diligence Act], July 16, 2021, BGBl. I S. 2959 (Germany). https://www.gesetze-im-internet.de/lksg/

7. Loi n° 2017-399 du 27 mars 2017 relative au devoir de vigilance des sociétés mères et des entreprises donneuses d'ordre [Duty of Vigilance Law], March 27, 2017 (France). https://www.legifrance.gouv.fr/loda/id/JORFTEXT000034290296

8. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation). *Official Journal of the European Union*, L 119, 1–88. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679

9. United Nations. (1948). *Universal Declaration of Human Rights.* https://www.un.org/en/about-us/universal-declaration-of-human-rights

10. United Nations. (1966). *International Covenant on Civil and Political Rights*, 999 U.N.T.S. 171. https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights

11. European Data Protection Board. (2023, May 22). *1.2 billion euro fine for Facebook as a result of EDPB binding decision.* https://www.edpb.europa.eu/news/news/2023/12-billion-euro-fine-facebook-result-edpb-binding-decision_en

12. Irish Data Protection Commission. (2023, May 22). *Data Protection Commission announces conclusion of inquiry into Meta Ireland.* https://www.dataprotection.ie/en/news-media/press-releases/Data-Protection-Commission-announces-conclusion-of-inquiry-into-Meta-Ireland

13. UN Human Rights Council. (2021). *Resolution 47/23: New and emerging digital technologies and human rights.* United Nations. https://www.ohchr.org/en/hr-bodies/hrc/regular-sessions/session47/res-dec-stat

Secondary Sources

Book

1. Bradford, A. (2020). *The Brussels effect: How the European Union rules the world.* Oxford University Press. https://global.oup.com/academic/product/the-brussels-effect-9780190088583

2. Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson. https://aima.cs.berkeley.edu

3. Voigt, P., & Von dem Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR): A practical guide.* Springer. https://doi.org/10.1007/978-3-319-57959-7

Journal Articles

1. Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

2. Calo, R. (2017). Artificial intelligence policy: A primer and roadmap. *UC Davis Law Review*, *51*(2), 399–435. https://lawreview.law.ucdavis.edu/issues/51/2/articles/files/51-2_Calo.pdf

3. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

4. Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

5. Raso, F., Hilligoss, H., Krishnamurthy, V., Bavitz, C., & Kim, L. (2018). *Artificial intelligence and human rights: Opportunities and risks.* Berkman Klein Center for Internet and Society. https://cyber.harvard.edu/publication/2018/artificial-intelligence-human-rights

6. Ruggie, J. (2011). *Guiding principles on business and human rights: Implementing the United Nations "Protect, Respect and Remedy" framework.* United Nations. https://www.ohchr.org/documents/publications/GuidingprinciplesBusinesshr_eN.pdf

7. Veale, M., & Borgesius, F. Z. (2021). Demystifying the Draft EU Artificial Intelligence Act. *Computer Law Review International*, *22*(4), 97–112. https://doi.org/10.9785/cri-2021-220402

8. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law and Technology*, *31*(2), 841–887. https://doi.org/10.2139/ssrn.3063289

Institutional and Policy Documents

1. Afina, Y. (2026, February 12). *Thirty-five nations back military AI rules as US, China opt out.* Mexico Business News. https://mexicobusiness.news/cloudanddata/news/thirty-five-nations-back-military-ai-rules-us-china-opt-out

2. Arms Control Association. (2024, August 8). *The Chemical Weapons Convention (CWC) at a glance.* https://www.armscontrol.org/factsheets/chemical-weapons-convention-cwc-glance-0
3. Caroli, L., & Mande, M. (2025, October 7). *What the UN Global Dialogue on AI Governance reveals about global power shifts.* Center for Strategic and International Studies. https://www.csis.org/analysis/what-un-global-dialogue-ai-governance-reveals-about-global-power-shifts

4. Dafoe, A. (2018). *AI governance: A research agenda.* Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf

5. Grand-Clément, S. (2023). *Artificial intelligence beyond weapons: Application and impact of AI in the military domain.* UNIDIR. https://unidir.org/publication/artificial-intelligence-beyond-weapons-application-and-impact-of-ai-in-the-military-domain/

6. He, W., & Anand, A. (2023). *The 2022 Innovations Dialogue: AI disruption, peace and security.* UNIDIR. https://unidir.org/publication/the-2022-innovations-dialogue-ai-disruption-peace-and-security-conference-report/

7. High-Level Expert Group on AI. (2019). *Ethics guidelines for trustworthy AI.* European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

8. OECD. (2019). *Recommendation of the Council on Artificial Intelligence.* Organisation for Economic Co-operation and Development.
9. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

10. OPCW. (n.d.). *History of the Chemical Weapons Convention.* Organisation for the Prohibition of Chemical Weapons. https://www.opcw.org/about-us/history

11. Puscas, I. (2023). *AI and international security: Understanding the risks and paving the path for confidence-building measures.* UNIDIR. https://unidir.org/publication/ai-and-international-security-understanding-the-risks-and-paving-the-path-for-confidence-building-measures/

12. Rapid Transition Alliance. (2019, June 11). *Back from the brink: How the world rapidly sealed a deal to save the ozone layer.* https://rapidtransition.org/stories/back-from-the-brink-how-the-world-rapidly-sealed-a-deal-to-save-the-ozone-layer/

13. UNEP. (n.d.). *The Montreal Protocol on substances that deplete the ozone layer.* United Nations Environment Programme Ozone Secretariat. https://ozone.unep.org

14. UNESCO. (2021, November). *Recommendation on the ethics of artificial intelligence.* United Nations Educational, Scientific and Cultural Organisation.https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

15. Vinals Musquera, A., & Babwah Brennen, S. (2026, April 20). *Regulatory uncertainty is what actually holds back innovation.* Brookings Institution.
16. https://www.brookings.edu/articles/regulatory-uncertainty-is-what-actually-holds-back-innovation/

17. Leo XIV. (2026, May 25). *Magnifica humanitas: On safeguarding the human person in the time of artificial intelligence* [Encyclical letter]. Vatican. https://www.vatican.va/content/leo-xiv/en/encyclicals/documents/20260515-magnifica-humanitas.html

18. ***News Sources***

19. Cadwalladr, C., & Graham-Harrison, E. (2018, March 17). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian.* https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

20. Hill, K. (2020, June 24). Wrongfully accused by an algorithm. *The New York Times.* https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

21. American Civil Liberties Union. (2021, April 13). *Michigan father sues Detroit Police Department for wrongful arrest based on faulty facial recognition technology.* https://www.aclu.org/press-releases/michigan-father-sues-detroit-police-department-wrongful-arrest-based-faulty-facial

22. Doniger, D. (2015, December 10). *Paris climate agreement explained: Does Congress need to sign off?* Natural Resources Defense Council. https://www.nrdc.org/bio/david-doniger/paris-climate-agreement-explained-does-congress-need-sign