

MenschenRechtsMagazin

Informationen | Meinungen | Analysen

30. Jahrgang · 2025 · Heft 2

Aus dem Inhalt

- Autonomous Weapon Systems and its Fundamental Flaw
- International Artificial Intelligence Law to the Test of Surveillance



Universitätsverlag Potsdam

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.dnb.de/> abrufbar.

Universitätsverlag Potsdam 2025

<http://verlag.ub.uni-potsdam.de/> | <https://ror.org/01femje42>

Am Neuen Palais 10, 14469 Potsdam

Tel.: +49 (0)331 977 2533

E-Mail: verlag@uni-potsdam.de

Herausgeber:

Prof. Dr. phil. Logi Gunnarsson (logi.gunnarsson@uni-potsdam.de)

Prof. Dr. iur. Andreas Zimmermann, LL.M. (Harvard) (andreas.zimmermann@uni-potsdam.de)

Begründet von: Prof. Dr. iur. Eckart Klein

Redaktion:

Prof. Dr. iur. Norman Weiß (norman.weiss@uni-potsdam.de)

Btissam Boulakhrif (redaktion-mrm@uni-potsdam.de)

Alle Rechte liegen bei den Autor:innen. Copyright-Jahr 2026.

Soweit nicht anders gekennzeichnet, ist dieses Werk unter einem Creative-Commons-Lizenzvertrag Namensnennung 4.0 lizenziert. Dies gilt nicht für Zitate und Werke, die aufgrund einer anderen Erlaubnis genutzt werden. Um die Bedingungen der Lizenz einzusehen, folgen Sie bitte dem Hyperlink:

<https://creativecommons.org/licenses/by/4.0/legalcode.de>.

Copyediting: Georgiy Kurie

Satz: le-tex publishing services GmbH, xerif



Online veröffentlicht unter:

<https://doi.org/10.60935/mrm2025.30.2>

Inhaltsverzeichnis

Norman Weiß, Btissam Boulakhrif Editorial	86
Beiträge	
Ruvini Katugaha Autonomous Weapon Systems and its Fundamental Flaw	87
William Letrone, Tony Cabus International Artificial Intelligence Law to the Test of Surveillance	97
Steven Kleemann, Milan Tahraoui Responsibility and Accountability for the Use of AI in Law Enforcement in the European Union – Lost in Negotiations?	117
Manuel Brunner The Impact of Artificial Intelligence on the Work of Human Rights Defenders	144

Editorial

Norman Weiß¹  • Btissam Boulakhrif¹ 

¹Universität Potsdam, MenschenRechtsZentrum

Dieses Heft widmet sich dem Themenkomplex der Künstlichen Intelligenz im Kontext des Menschenrechtsschutzes aus unterschiedlichen Perspektiven. Die Beiträge bauen dabei auf Vorträgen auf, die im Rahmen der Tagung *Human Rights and Artificial Intelligence: Addressing Challenges, Enabling Rights* vom 7. bis 8. November, anlässlich des 30-jährigen Jubiläums des MenschenRechtsZentrums, präsentiert wurden.

Ruvini Katugaha befasst sich in ihrem Beitrag „Autonomous Weapon Systems and its Fundamental Flaw“ mit der Frage der Regulierbarkeit autonomer Waffensysteme. Ausgehend von deren Auswirkungen auf ausgewählte menschenrechtliche Gewährleistungen sowie bestehenden Regulierungsansätzen arbeitet sie Herausforderungen und Hürden der Regulation heraus.

William Letrone und *Tony Cabus* analysieren in ihrem Beitrag „International artificial intelligence law to the test of surveillance“ die rechtlichen Rahmenbedingungen Künstlicher Intelligenz im Bereich des Datenschutzes. Sie befassen sich hierfür zunächst mit dem Einsatz Künstlicher Intelligenz im Bereich privater und staatlicher Überwachung sowie mit dem Konzept digitaler Privatsphäre. Auf dieser Grundlage untersuchen sie, inwieweit bestehende rechtliche Regelungen einen angemessenen Schutz digitaler Privatsphäre gewährleisten.

In ihrem Beitrag „Responsibility and Accountability for the use of AI in Law Enforcement in the European Union: Lost in Negotiations?“ untersuchen *Steven Kleemann* und *Milan Tahraoui* die Anwendung der KI-Verordnung der Europäischen Union im Rahmen der Strafverfolgung. Hierfür analysieren sie den risikobasierten Ansatz der Regulation im Rahmen der Strafverfolgung sowie die Rechenschafts- und Anfechtungsmechanismen in Bezug auf die Nutzung künstlicher Intelligenz.

Manuel Brunner diskutiert schließlich die Implikationen künstlicher Intelligenz auf den Schutz sowie die Arbeit von Menschenrechtsverteidigerinnen in seinem Beitrag „The Impact of Artificial Intelligence on the Work of Human Rights Defenders“.

Die freigewordene Stelle im Redaktionsteam konnte das MRZ zum 1. Oktober 2025 mit Frau Btissam Boulakhrif besetzen; sie hat die Entstehung dieser Ausgabe komplett begleitet. Unseren Leser:innen wünschen wir eine erkenntnisreiche sowie anregende Lektüre.

Autonomous Weapon Systems and its Fundamental Flaw

Ruvini Katugaha¹

¹University of the West of England

Contents

- I. Introduction
 - II. The current debate on AWS
 - III. The impact of the use of AWS on human rights
 - IV. The current approach to regulation
 - V. The challenge of regulating fully AWS
 - VI. Conclusion
- Vita

Abstract

Highly autonomous weapons can make split-second decisions about life and death without any human involvement thereby avoiding human accountability in the decision-making process. Accountability is an essential component for the proper functioning of the law. All law is premised on human agency. Thus, human agency is essential to accountability. The lack of human agency poses a challenge to the regulation of artificial intelligence.

Using Autonomous Weapon Systems (AWS) as an example, this research paper will explore the challenge of regulating highly “intelligent” and “autonomous” AI-incorporated weapons using a sociolegal methodology employing doctrinal, theoretical and comparative methods of research. While incorporating AI into weapons is not inherently harmful, the paper concludes that it is impossible to regulate “fully” AWS (which incorporates sophisticated AI into the weapon system) because human agency is absent in the “decision” to apply “lethal force”, which undermines accountability. Furthermore, even when human involvement is present, it occurs at different stages of the process and does not necessarily include the decision-making phase. Thus, it is submitted that a fully AWS carries with it the fundamental flaw that it cannot be regulated by law.

Citation:

Ruvini Katugaha, Autonomous Weapon Systems and its Fundamental Flaw, in: MRM 30 (2025) 2, pp. 87–96.
<https://doi.org/10.60935/mrm2025.30.2.30>.

Received: 2025-09-02

Accepted: 2026-01-05

Published: 2026-02-17

Keywords

Autonomous Weapon Systems, Artificial Intelligence, Human Rights, Accountability, Regulation

Permissions:

The copyright remains with the authors.
Copyright year 2026.

Unless otherwise indicated, this work is licensed under a [Creative Commons License Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/). This does not apply to quoted content and works based on other permissions.

I. Introduction*

Artificial intelligence (AI) has come a long way in the last decade. Its integration in our day-to-day lives has reached a point where AI has an impact on human rights. Academics are widely divided on the impact AI has on human rights, though all acknowledge its benefits and challenges. Yet, all contend that, as with everything else in our daily lives, AI-related technology does require *some* form of regulation. The key question is, however, *whether it can be regulated*. While most AI-related technology can be regulated (e.g., during manufacturing, procurement, use, etc.), there appears to be an emerging category of AI-incorporated technology that may prove difficult, if not impossible, to regulate due to its nature. One such category is the incorporation of AI in the context of weapons, particularly Autonomous Weapon Systems (AWS).

There is no universally agreed-upon definition of AWS, as States disagree on what constitutes an AWS. For this research paper, the approach will be to outline the elements that constitute an AWS rather than relying on an existing definition. These elements are: a high level of autonomy, a lack of or minimal human control, the ability to make lethal decisions, and the unpredictability of those decisions from a human perspective.¹ Thus, AWS

are weapon systems capable of acting autonomously with minimal or no human control in the lethal decision-making process. They mainly include two types: semi-autonomous weapon systems (SAWS) and fully autonomous weapon systems (AWS). SAWS are weapon systems that maintain some form of meaningful human control in the lethal decision-making process, whereas fully AWS lack such control. The reason for emphasizing meaningful human control is that a lack thereof leads to a lack of human judgment in the decision to employ lethal force. This paper focuses solely on fully AWS, as almost all of them incorporate AI, thereby granting the weapon system the capacity to be ‘highly intelligent’ (depending on the type of AI) and thus allowing it to be deemed an ‘intelligent’ weapon system. In this paper, ‘intelligent’ refers to AI that enables the weapon to perform certain high-level cognitive functions (typically performed by humans) without human involvement or control. It should be emphasized that there is a growing trend in most AWS (though not all) to incorporate machine learning, allowing the AWS not only to operate independently from a human but also to ‘think’ and act independently or autonomously and ‘learn’ on its own. AI incorporated into everyday objects like cars and phones may have positive implications, and using AI for military purposes might not necessarily be negative *per se*. It is important to note that this paper does not take the position that incorporating AI into weapon systems is inherently negative, nor does it argue that the use of highly ‘intelligent’ AWS is detrimental. As *Scharre* states, “many military applications of AI are uncontroversial—improved logistics, cyber defences, and robots for medical evacuation, resupply, or surveillance—however,

* This paper was presented at the 30th Anniversary Conference on Human Rights and Artificial Intelligence: Addressing challenges, enabling rights held on the 7th and 8th of November 2024 in Potsdam, Germany.

¹ Please note that this definition, extracted from the components that constitute an AWS, is drawn from my PhD thesis, which is yet to be published. The components were identified by extensively analysing definitions adopted by States (submitted to the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Au-

tonomous Weapons System), international and regional organizations, NGOs and legal scholars.

the introduction of AI into weapons raises challenging questions”.² One challenge is that as AI becomes more sophisticated (or ‘intelligent’), it also becomes more autonomous. While greater autonomy may seem acceptable and beneficial in most cases, it could be disastrous on a battlefield.

Therefore, the key question is not only how AI-integrated fully AWS can be regulated but, more fundamentally, *whether* such weapons can be effectively regulated at all. The need for regulation arises because the AI-integrated fully AWS, unlike other weapon systems or pieces of technology, can independently decide to apply lethal force against humans. It is argued that, without proper regulation, this capability could lead to arbitrary loss of lives (thereby violating Human Rights Law and International Humanitarian Law (IHL)). While IHL remains *lex specialis* in armed conflicts, Human Rights Law applies concurrently.³ Accordingly, this paper focuses on the use of fully AWS and its impact on human rights in the context of an armed conflict.⁴ Furthermore, it will essentially, though not exclusively, focus on non-derogable human rights, as other human rights

can be derogated from, subject to conditions,⁵ during an armed conflict.

This paper explores the impact of AI-integrated fully autonomous weapons on human rights and draws the conclusion that incorporating such AI into weapons systems (in a manner that makes that weapon or weapon system a ‘fully’ AWS) makes it, by nature, impossible to regulate. This is because human agency is absent in the ‘decision’ to apply ‘lethal force’, thereby eliminating the element of accountability.

This paper will first discuss the current debate on AWS by briefly highlighting the arguments brought forth by those who are for and against the use of AWS. Secondly, it will discuss the impact of using AWS on some selected human rights. Thirdly, it will bring to focus the current positions on regulating AWS. Then it will focus on the challenges to regulating AWS, exposing its fundamental flaw. The paper will conclude that fully AWS have a fundamental flaw, making them unable to be regulated by law.

II. The current debate on AWS

The current debate on AWS warrants a brief discussion to highlight the broad spectrum of opinions regarding its regulation and notably that of fully AWS.⁶ Those opposed to the use and development of

² Paul Scharre, *Army of None: Autonomous Weapons and the Future of War*, 2018, p. 12.

³ UNHRC, General Comment No. 31: Nature of the General Legal Obligation Imposed on States Parties to the Covenant of 26 May 2004, UN Doc CCPR/C/74/CRP.4/Rev.6./C/GC/35, para. 11; *Legal Consequences of the Construction of a Wall in the Occupied Palestinian Territory*, Advisory Opinion, ICJ Reports (2004), pp. 136, paras. 101 ff. and 127 ff.

⁴ Thus, the use of AWS in law enforcement is not covered. As it stands now, the technology of the AWS can be used to attack demonstrators but have not been used as such by any State (which could potentially occur in the distant future).

⁵ International Covenant on Civil and Political Rights of 16 December 1966, UNTS vol. 999, p. 171, (ICCPR) Art 4.

⁶ It must be noted that when different States use the term ‘AWS’ some include both SAWS and fully AWS. Others simply equate AWS with fully AWS.

AWS⁷ fear that humans will delegate the decision to use lethal force (thereby transferring the power to take lives) into the hands of a weapon system, which has no feelings or remorse.⁸ In a sense, they worry that these systems will be unable to exercise human judgment in the battlefield in determining what is right or wrong, or what is lawful or not. This lack of exercise of human judgement may lead to a serious breakdown of the law itself, as AWS cannot be held accountable for their actions. The foundation of law rests on the principle that those governed by it will adhere to the parameters and be held accountable for failing to do so.

The opposing view is that a complete prohibition on fully AWS cannot be achieved. Several reasons⁹ are evoked to support this stance, including the right to self-defence as outlined in the UN Charter.¹⁰ These proponents argue that AI-driven AWS should be available for use in self-defence in the event of an attack with similar weapons. They believe that only AI-driven weapons can successfully counter other AI-driven weapon systems. They argue that banning AWS would leave them vulnerable, unable to defend themselves. Moreover, they argue that employing AWS in the battlefield could significantly reduce the loss

of combatant and civilian lives, improve objectivity and accuracy, and even wage war ethically.¹¹

III. The impact of the use of AWS on human rights

While some may argue that the use of AWS falls exclusively in the domain of IHL as it is *lex specialis*,¹² the human rights law perspective cannot be ignored. For example, the 2013 Report of the UN Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions highlighted the importance of regulating AWS.¹³ Moreover, regional collectives, such as the Belén Communiqué of the Latin American and the Caribbean Conference of Social and Humanitarian Impact of Autonomous Weapons,¹⁴ and the Caribbean Community (CARICOM) Declaration on Autonomous Weapons Systems at the CARICOM Conference,¹⁵ have recognized the impact of the deployment of AWS on human rights.

⁷ An interesting point to note is that many who are opposed to the development and use of AWS are, in fact, against ‘fully’ AWS. Thus, when they call for a ban on AWS, what they actually mean is fully AWS. The confusion in the terminology has indeed not helped matters.

⁸ *Bonnie Docherty*, *Losing Humanity: The Case against Killer Robots*, 2012, p. 4.

⁹ See for the various arguments supporting the use of AWS: *Christopher P. Toscano*, *Friend of Humans: An Argument for Developing Autonomous Weapons Systems*, in: *Journal of National Security Law and Policy* 8 (2015), pp. 189–246.

¹⁰ Charter of the United Nations of 26 June 1945, UNTS vol. 1, p. XVI (UN Charter).

¹¹ *Ronald C. Arkin*, *The Case for Ethical Autonomy in Unmanned Systems*, in: *Journal of Military Ethics* 9 (2010), pp. 332–341.

¹² *Toscano* (fn. 9), p. 50.

¹³ UNHRC, Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions of 9 April 2013, UN Doc A/HRC/23/47.

¹⁴ Latin American and the Caribbean Conference of Social and Humanitarian Impact of Autonomous Weapons, *The Belén Communiqué* of 24 February 2023, reproduced in: United Nations General Assembly, *Lethal Autonomous Weapons Systems - Report of the Secretary-General*, UN Doc A/79/88 of 1 July 2024, pp. 40–41.

¹⁵ CARICOM, *Declaration on Autonomous Weapons Systems*, CARICOM Conference: *The Human Impacts of Autonomous Weapons*, Port of Spain Trinidad and Tobago, 5–6 September 2023, available at: https://www.caricom-aws2023.com/_files/ugd/b69acc_c1ffb97ed9024930a3205ae4e34c1b45.pdf (last visited 17 December 2025).

To illustrate the impact of the use of AWS on human rights, this paper examines the concept of human dignity, which is fundamental to all human rights, the principle of non-discrimination that applies in concurrence with all human rights, and three specific human rights: the right to life, the prohibition of torture, and the right to privacy.

1. Human dignity

Every single human right is based on the underlying concept of human dignity.¹⁶ As the Universal Declaration on Human Rights (UDHR) emphasizes in its preamble,¹⁷ the “recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world”. Human rights are inherent in every human being by virtue of simply being human, whether they are criminals or law-abiding citizens. Again, Art. 1 of the UDHR reminds us that all “human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood”.

Undoubtedly, human dignity, the building block of all other human rights, is seriously affected in times of armed conflict. Granting AWS the power to make life-and-death decisions in armed conflict presents a fundamental challenge to the concept of human dignity. First, these weapons lack the capacity to comprehend and respect human dignity. To AWS, a human

or a human combatant is merely digits and numbers in a mechanical process of calculating and selecting whether to kill. This, in a way, dehumanizes and reduces the value of a human being. As a Human Rights Watch report clearly points out fully AWS are inanimate machines, they “could truly comprehend neither the value of individual life nor the significance of its loss. Allowing them to make determinations to take life away would thus conflict with the principle of dignity”.¹⁸ Whilst it is true that IHL mandates the categorisation of people under the principle of distinction, thus reducing humans to legitimate (e.g. combatants) and unlawful (e.g. civilians) targets, one of the key principles of IHL is also the principle of humanity. For example, wounded or incapacitated combatants, as well as those who surrender, cannot be targeted. But how would the AWS understand that the combatant is unwell beyond its physical appearance? Would it be able to show compassion and then treat the combatant turned *hors de combat* in a humane manner?

Second, in the words of the UN Special Rapporteur on extra-judicial killing, the use of AWS means that “in addition to being physically removed from the kinetic action, humans would also become more detached from decisions to kill – and their execution”.¹⁹ In turn, this detachment may lead to an increased willingness to kill ‘the enemy’ (i.e. launch an AI-integrated AWS) without the burden of conscience or accountability. The premise is that it is only a human who can recognize the intrinsic value of another human.²⁰ Ma-

¹⁶ UN Charter, preamble; Universal Declaration of Human Rights of 10 December 1948, UN Doc. A/RES/217 A (III), preamble; ICCPR, preamble; International Covenant on Economic, Social and Cultural Rights of 16 December 1966, UNTS vol. 993, p. 3 (ICESCR), preamble.

¹⁷ See also the preamble of the ICCPR.

¹⁸ *Bonnie Docherty*, *Shaking the Foundations: The Human Rights Implications of Killer Robots*, 2014, p. 3.

¹⁹ UNHRC (fn. 13), para. 27.

²⁰ After World War II, human dignity was conceived as a prerequisite for human coexistence and sol-

chines, at their current stage of technological development, cannot fathom the concept of human dignity or the value of human life. Therefore, as it is impossible to train machines to ‘value’ human lives, AI-integrated AWS cannot be reliably regulated to ensure they respect human dignity.

2. The principle of non-discrimination

Like the concept of human dignity, the principle of non-discrimination is fundamental in human rights law. The ICCPR, like many human rights treaties and declarations,²¹ emphasizes in its Art. 27 that “the law shall prohibit any discrimination and guarantee to all persons equal and effective protection against discrimination on any ground such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status”. While some may argue that an AWS does not have biases like humans do (as it is just a machine and is therefore neutral), all algorithms will reflect the biases of the programmers or parties manufacturing the AI or weapon system. This may lead to racial profiling or discrimination based on various other grounds, sometimes very subtly weaponizing racism or misogyny. It is thus difficult, if not impossible, to ensure that a fully AWS does not violate the principle of non-discrimination under human rights Law.

3. The right to life

The right to life is a basic and fundamental right that is non-derogable. According

identity (see the use of the terminology such as “members of the human family” and “brotherhood” in the UDHR) and thus can only be understood as a concept that works between humans. It carries with it an idea of shared humanity.

²¹ ICESCR, Art. 2 para. 2; UDHR, Art. 7.

to Article 6 of the ICCPR, “every human being has the inherent right to life (...) No one shall be arbitrarily deprived of his life”. This is reaffirmed in other treaties and declarations as well.²² The Human Rights Committee (UNHRC) noted in General Comment No. 36 (2019) that this right is “the supreme right from which no derogation is permitted, even in situations of armed conflict or other public emergencies that threaten the life of the nation”.²³

As early as 2013 the attention of the Human Rights Council was drawn to the impact of AWS. As the Special Rapporteur’s report pointed out “the introduction of such powerful yet controversial new weapons systems has the potential to pose new threats to the right to life”.²⁴ The horror of humans taking the lives of humans is now further complicated by the arrival of AWS which lack empathy or forgiveness. As the Special Rapporteur correctly points out:

“One of the most difficult issues that the legal, moral and religious codes of the world have grappled with is the killing of one human being by another. The prospect of a future in which fully autonomous robots²⁵ could exercise the power of life and death over human beings raises a host of additional concerns.”²⁶

From a legal viewpoint, the problem with, especially fully, AWS is that, in warfare, they might not be able to comply with the prohibition of *arbitrary* killing. Indeed, according to human rights law, not every

²² UDHR, Art. 3.

²³ UNHRC, General Comment No. 36: Article 6 (Right to Life) of 3 September 2019, UN Doc. CCPR/C/GC/35, para. 2.

²⁴ UNHRC (fn. 13), para. 30.

²⁵ He uses the term robots as a similar word to AWS.

²⁶ UNHRC (fn. 13), para. 30.

taking of human life is prohibited; only that which is deemed arbitrary.²⁷ When fully AWS engage in warfare, they essentially make decisions about using force without human intervention. The decisions made by AWS in any given situation can be extremely unpredictable for the humans who authorized the use or launched the weapon. This unpredictability can result in arbitrary killings that cannot be avoided as humans are taken out of the decision-making loop.

4. Prohibition of torture

The use of AWS also raises concerns regarding the prohibition of torture, which is a non-derogable right enshrined in Article 7 of the ICCPR²⁸ in the following manner: “no one shall be subjected to torture or to cruel, inhuman or degrading treatment or punishment”. There are two scenarios in which the use of AWS could lead to a violation of the prohibition of torture and other forms of inhuman treatment. First, an AWS, programmed to kill, might fail to do so and, instead, inflict immense pain to humans. Second, depending on how they are programmed and manufactured, AWS could potentially be used with the intention to inflict torture or other types of inhuman treatment. It would be exceedingly difficult for a weapon or system to discern what constitutes ‘degrading’ or ‘cruel’ treatment, especially in context-specific situations. Even if AWS were able to gather data on the health of a human, they would likely be unable to assess physical and mental pain accurately and, thus, would struggle to determine whether they

are engaging in torture or cruel, inhuman or degrading treatment or punishment.

5. The right to privacy

Although the right to privacy in Art. 8 IC-CPR which prohibits “arbitrary or unlawful interference with his privacy, family, home or correspondence [...]” is not a non-derogable right, its examination is warranted because it highlights the complex nature of using fully AWS. In fact, an AI-integrated fully AWS might be able to tap into a variety of platforms and databases, mining for information. In today’s technological era, people’s information, such as medical records, national identity cards, and genetic data, is often stored in government and private databases. This information can be accessed and used for profiling individuals as well as engaging in targeted killings. If granted such access, AWS could easily profile individuals in an armed conflict and deploy force against them. Additionally, there is a risk that such data falls into the hands of armed non-State actors as well as other States (in times of occupation).

To summarize my position so far, fully AWS (that integrates high-level AI technology) would not be able to comply with human rights law.

IV. The current approach to regulation

Currently, there are three different approaches put forward by States, international organisations and NGOs on the matter of regulating AWS, namely:

²⁷ For an example, under the General Comment No. 36, the use of lethal force in self-defence does not constitute an arbitrary deprivation of life. See UNHRC (fn. 23), para. 10.

²⁸ See also Art. 5 of UDHR

- i. Total prohibition of AWS (what many States mean by AWS here is fully AWS).
- ii. Regulation of AWS (without a complete prohibition) through a treaty or a non-legally binding code of conduct.
- iii. Regulation through a two-tier approach of prohibiting fully AWS without any meaningful human control and regulating those with some form of meaningful human control.

Advocates of a ban on fully AWS, such as the Campaign to Stop Killer Robots (including Human Rights Watch) and States such as Canada, believe that fully automated AI-driven AWS could not ever be used in a manner compliant with IHL. They are of the opinion that the “use of an AWS whose operation, behaviour and effects cannot be limited according to IHL, notably the principles of distinction, proportionality and precaution, would be unlawful”.²⁹ Many NGOs and human rights organizations, such as those that are part of the Stop Killer Robots Campaign, find that “fully autonomous weapons would not only be unable to meet legal standards but would also undermine essential non-legal safeguards for civilians.”³⁰ Those supporting the use of AWS³¹ conclude that

a categorical prohibition on AWS is unjustified. States, such as the United States of America, advocate for the development of a non-binding code of conduct on the matter.

It must be stressed that States approach regulation from an IHL perspective. As seen in the above-mentioned approaches, since the use of fully AWS has emerged in the context of armed conflict, States automatically consider addressing the matter through IHL. However, they overlook the fact that there is a human rights dimension to the use of fully AWS in armed conflict, as human rights law also applies in times of armed conflict and because AWS can also violate human rights law. Thus, human rights-related approaches can contribute to improving and guiding the understanding of the debate surrounding the use and development of fully AWS.

The most preferred form of regulation appears to be the two-tier approach.³² Most States support the ban of fully AWS with no meaningful human control while advocating for a set of rules (preferably through an IHL treaty) for SAWS or non-fully AWS. However, States already engaged in the development and research of AWS (most importantly fully AWS) have managed to evade responsibility because there is no consensus on what constitutes fully AWS. They can sidestep liability by simply adopting a higher threshold for the definition of fully AWS, thus arguing that these weapons are not fully AWS and thus not

²⁹ Vincent Boulanin/Netta Goussac/Laura Bruun, *Autonomous Weapon Systems and International Humanitarian Law: Identifying Limits and the Required Type and Degree of Human-Machine Interaction*, SIPRI of June 2021, available at: <https://www.sipri.org/publications/2021/policy-reports/autonomous-weapon-systems-and-international-humanitarian-law-identifying-limits-and-required-type> (last visited 20 September 2024).

³⁰ Docherty (fn. 8).

³¹ Toscano (fn. 9); Kenneth Anderson/Matthew C. Waxman, *Debating Autonomous Weapon Systems, their Ethics, and their Regulation under International Law*, in: Roger Brownsword/Eloise Scotford/Karen Yeung (ed.), *The Oxford Hand-*

book of Law, Regulation and Technology, 2017, pp. 1097-1117.

³² Supported by States such as Austria and China. See United Nations Office for Disarmament Affairs, *The United Nations Disarmament Yearbook 2023*, Vol. 48 (2024), available at: <https://media-publications.unoda.org/documents/full-en-yb-vol-48-2023.pdf> (last visited 17 December 2025), p. 124.

banned. The two-tier approach might thus not be effective unless an international agreement is reached on what constitutes a fully AWS.

V. The challenge of regulating fully AWS

One of the fundamental challenges of fully AWS is the lack of accountability. As the Belén Communiqué of the Latin American and the Caribbean Conference of Social and Humanitarian Impact of Autonomous Weapons points out “it is paramount to maintain meaningful human control to prevent further dehumanization of warfare, as well as to ensure individual accountability and state responsibility”.³³

According to international law (to which the human rights law regime belongs), the responsibility of a State is engaged when there is an internationally wrongful act that 1) “is attributable to the State” and 2) “constitutes a breach of an international legal obligation”.³⁴ As military personnel are State agents,³⁵ and thus organs of a State according to Article 4 of the Draft Articles on the Responsibility of States for Internationally Wrongful Acts, their actions are to be scrutinized with a view to determining whether the State has violated human rights law. Yet, when it involves fully AWS,

holding the State accountable for human rights breaches can be challenging. In fact, it may only be possible to hold the agent accountable for the launch of the AWS, not for its subsequent actions. At the time of the launch, the agent may have assessed the situation and deemed that deploying the AWS would comply with human rights law. Nonetheless, once deployed, the AWS may have acted in contravention of human rights law.

The key concepts in this context are human agency and human judgment. All humans possess human agency, which is the capacity to make decisions and act on them. Human judgement, comprised of moral, ethical, and legal building blocks, is a fundamental aspect of that human agency as it guides humans in exercising their agency in a morally, ethically, and legally appropriate manner. A person cannot be held accountable for something they could not predict or perceive at the time of the action. In fully AWS, there is no meaningful human control beyond the launch of the system, and so State agents can only be held responsible for that *launch* when they were exercising their human agency as constrained by their human judgement.

The Special Rapporteur sums up my argument well in the following words:

“Armed conflict and IHL often require human judgement, common sense, appreciation of the larger picture, understanding of the intentions behind people’s actions, and understanding of values and anticipation of the direction in which events are unfolding. Decisions over life and death in armed conflict may require compassion and intuition. Humans – while they are fallible – at least might possess these qualities, whereas robots definitely do not...[T]hey have limited abilities to make the qualitative assessments that are often called for when dealing with human life. Machine calculations

³³ Latin American and the Caribbean Conference of Social and Humanitarian Impact of Autonomous Weapons (fn. 14).

³⁴ See UN General Assembly, Responsibility of States for internationally wrongful acts, UN Doc. A/RES/56/83 of 28 January 2002, Annex, Arts. 1 and 2.

³⁵ Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I) of 8 June 1977, UNTS vol. 1125, p. 3, Art. 91.

are rendered difficult by some of the contradictions often underlying battlefield choices.”³⁶

The absence of any human involvement is a distinguishing feature of fully AWS. While States may argue that there is human involvement in the various stages of the AWS (such as in programming, manufacturing, and authorizing the use and deployment) the defining feature of fully AWS is the complete absence of meaningful human involvement or control in the final *decision* to use lethal force. Therefore, by nature, such fully AWS make life-and-death decisions without any human input. Here lies its fundamental flaw. When meaningful human involvement is eliminated from the equation, it automatically removes human agency and, thereby, human judgment. The lack of human judgment makes it impossible to hold anyone accountable for the actions of fully AWS. Indeed, the State agent may deny any responsibility for the violation of international law; although it was the agent who launched the weapon, they are not held accountable for the attack and the ensuing death of civilians, for example. The State can thus deny responsibility on the basis that the act of the AWS cannot be attributed to it; after all, the agent could not predict what the AWS would do. Consequently, it is contended that such AWS cannot, under the current legal framework, be effectively regulated.

VI. Conclusion

In conclusion, the complexities surrounding the regulation of lethal force by AI-integrated fully AWS necessitates a more

nuanced understanding of both technological capabilities and legal implications. The absence of human agency involved in the final decision to use lethal force presents significant regulatory challenges. Even if there is some form of human involvement present, such involvement materializes itself at different stages and not in the final decision to use lethal force as such. Thus, it is submitted that AWS carry with them the fundamental flaw that they cannot be regulated by law.

Indeed, traditional frameworks that require human oversight and thus responsibility will struggle to address the rapid advancements of such technology. As conflicts become increasingly influenced by technology, the debate surrounding the compliance of fully AWS with human rights law cannot be ignored. Those advocating for stricter regulations will have to grapple with the potential for misuse, accidental engagement and the legal implications of delegating life-and-death decisions to machines. Moreover, the challenge of establishing accountability in situations where AWS operate in a fully autonomous mode raises further questions about State responsibility. To move towards (effective) regulation, it is crucial to explore innovative frameworks that encompass the unique characteristics of AWS.

Vita

The author is a doctoral student at the University of the West of England, UK, with a research interest in International Humanitarian Law. Her Ph.D. research is focused on regulating autonomous weapon systems in armed conflicts.

³⁶ UNHRC (fn. 13), para. 55.

International Artificial Intelligence Law to the Test of Surveillance

William Letrone¹  • Tony Cabus² 

¹DCS, Nantes University, CNRS

²Walther-Schücking Institute for International Law, Kiel

Contents

- I. Introduction
 - II. AI systems for digital surveillance
 - III. Beyond data protection; making sense of privacy in the digital era
 - IV. AI-driven surveillance in emerging AI laws
 - V. Conclusive remarks on advancing AI privacy discussions
- Vitae

Abstract

This paper takes a broad look at the privacy implications of emerging supranational frameworks on artificial intelligence (AI), taking AI-driven surveillance by the private and public sectors as a case example of privacy-adverse practices. To do so, this paper first examines the relationship between AI technologies and surveillance practices, highlighting the privacy risks raised by corporate surveillance and state surveillance. The paper then recalls the scope and content of privacy, before pinpointing remaining gaps in emerging frameworks on AI that stand in the way of achieving robust privacy guarantees in the context of AI-driven surveillance.

Keywords

Privacy, Surveillance, Artificial intelligence, International law, Digital law

Citation:

William Letrone/Tony Cabus, International Artificial Intelligence Law to the Test of Surveillance, in: MRM 30 (2025) 2, pp. 97–116.
<https://doi.org/10.60935/mrm2025.30.2.26>.

Received: 2025-07-04

Accepted: 2025-11-21

Published: 2026-02-17

Permissions:

The copyright remains with the authors.
Copyright year 2026.

Unless otherwise indicated, this work is licensed under a [Creative Commons License Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/). This does not apply to quoted content and works based on other permissions.

“Man has become a document like any other, with an identity that he no longer ‘owns’, over which he has little control (...) and whose commercial purpose he underestimates.”^{*1}

I. Introduction

What are the privacy implications of emerging supranational frameworks on artificial intelligence (hereinafter AI)²? Especially since the release of generative AI, there have been multiple calls to action for addressing the risks posed by the deployment and use of AI systems. In many instances, these calls were followed by non-binding initiatives and technology-specific frameworks intended to complement existing frameworks such as data-protection regulations.

Initiatives such as the European Union’s latest Artificial Intelligence Act (AI Act)³ and the Council of Europe’s Framework Convention on AI (Framework Convention) are most welcome. The other option, self-regulation by the tech sector, mostly through soft law, is not desirable due to the underlying economic interests driving their activities. Binding rules are preferable to soft law instruments, although the speed at which the sector is evolving requires that caution be exercised throughout any legislative processes taken to this end. Furthermore, many AI-driven activities know no border. Regulating technologies of such transnational nature requires concerted regulatory efforts at the global level.⁴ Hence, the ongoing “rush to AI regulation”⁵ is an opportunity to collectively address longstanding privacy issues in light of technological advances in the AI domain.

Among all possible uses of AI technology, surveillance activities, as the act of “watching, listening to, or recording of an

* This paper was presented at the MenschenRechtsZentrum’s 30th Anniversary Conference “Human Rights and Artificial Intelligence – Addressing challenges, enabling rights,” of 7th/8th November 2024, at the panel “AI as a challenge to regulation.”

¹ [Translated from French by the authors]: « *L’Homme est devenu un document comme les autres, disposant d’une identité dont il n’est plus ‘propriétaire’ dont il ne contrôle que peu la visibilité (...) et dont il sous-estime la finalité marchande.* » Ertzscheid, O., *L’homme, un document comme les autres*, in : Hermès 53 (2009), pp. 33-40 (38).

² This work is based on the definition of AI system contained in the Council of Europe’s Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (Framework Convention), which is of international reach. Hence, the Convention defines AI as “a machine-based system that for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations or decisions that may influence physical or virtual environments.” See Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law of 05 September 2024, CETS No. 225 (Framework Convention), Art. 2.

³ EU Regulation 2024/1689 of 12 July 2024, OJ L, 2024/1689 (AI Act).

⁴ Talita de Souza Dias/ Rashmin Sagoo, *AI Governance in the Age of Uncertainty: International Law as a Starting Point*, Just Security of 2 January 2024, available at: <https://www.justsecurity.org/90903/ai-governance-in-the-age-of-uncertainty-international-law-as-a-starting-point/> (last visited 16 October 2025).

⁵ Expression borrowed from Nathalie Smuha, Biden, Bletchley, and the emerging international law of AI, *Verfassungsblog* of 15 November 2023, available at: <https://verfassungsblog.de/biden-bletchley-and-the-emerging-international-law-of-ai/> (last visited 11 November 2025) See also Itsiq Benizri/Arianna Evers/Shanon Togawa Mercer/Ali A. Jessani, *A Comparative Perspective on AI Regulation*, *Lawfare* of 17 July 2023, available at: <https://www.lawfaremedia.org/article/a-comparative-perspective-on-ai-regulation> (last visited 11 November 2025).

individual's activities"⁶, constitute major sources of privacy erosion. By definition, surveillance is antithetical to privacy. Regulating corporate surveillance and state surveillance would thus go a long way in order to mitigate many privacy risks stemming from AI technology. Moreover, United Nations members have expressly called "upon all Member States and, where applicable, other stakeholders to refrain from or cease the use of artificial intelligence systems that are impossible to operate in compliance with international human rights law or that pose undue risks to the enjoyment of human rights."⁷ Yet, no other area encapsulates the complexities and challenges of AI technology as profoundly as the surveillance practices of the private and public sectors, since these may involve AI tools at different levels. Although the burgeoning regulatory landscape pertaining to AI reveals that legislators worldwide have identified the key challenges associated with the technology, privacy generally takes the backseat.

This paper seeks to assess how some privacy concerns raised by AI - and their underlying causes - are actually being addressed in emerging supranational legal frameworks on AI, focussing on AI-driven surveillance. To this end, the paper starts by exploring how AI technology and digital surveillance practices intersect (II). In a second section, the paper offers an attempt to conceptualise digital privacy (III). The paper then analyses emerging AI regulations, highlighting the remaining gaps when it comes to mitigating AI-driven surveillance (IV). A few concluding remarks question the capacity of interna-

tional human rights law to rescue privacy (V).

II. AI systems for digital surveillance

Public and private actors are increasingly resorting to AI in their surveillance apparatus.⁸ This part examines AI uses in the context of digital surveillance, starting with digital surveillance by the private sector, or "corporate surveillance" (1), before moving to digital surveillance by the public sector, or "state surveillance" (2).

1. AI in corporate surveillance

"Corporate surveillance" is a synonym of "surveillance capitalism", a term famously coined by Harvard Professor *emerita* Shoshana Zuboff, which she defined as "the unilateral claiming of private human experience as free raw material for translation into behavioural data."⁹

"Surveillance capitalism" thus refers to a paradigm where individuals' behaviours are tracked, their desires inferred and anticipated based on the information collected from them, for the purpose of steering consumption habits. When describing

⁶ Daniel J. Solove, A Taxonomy of Privacy, in: University of Pennsylvania Law Review 154 (2006), pp. 477-560 (490).

⁷ UN Doc. A/RES/78/265, para. 5.

⁸ Steven Feldstein, The Global Expansion of AI Surveillance, Working Paper, Carnegie Endowment for International Peace, 2019, p. 6.

⁹ John Laidler, High tech is watching you, The Harvard Gazette of 4 March 2019, available at: <https://news.harvard.edu/gazette/story/2019/03/harvard-professor-says-surveillance-capitalism-is-undermining-democracy/> (last visited 11 November 2025); See also Joseph Jones, Don't Fear Artificial Intelligence, Question the Business Model: How Surveillance Capitalists Use Media to Invade Privacy, Disrupt Moral Autonomy, and Harm Democracy, in: Journal of Communication Inquiry 49 (2024), pp. 6-26 (9).

the advent of surveillance capitalism, authors speak of the “commodification” of both data and attention. Corporate surveillance proceeds from the dehumanizing rationale that economic value can (and must) be attached respectively, to users’ data and attention.¹⁰ While the underlying logics of corporate surveillance were already present in the advertising industry,¹¹ the ubiquity of AI surveillance tools makes it a more concerning trend today.

To commit corporate surveillance, digital companies rely on vast amounts of information *i.e.* big data, which aggregates information from a variety of sources.¹² This data is then exploited by algorithms to derive new insights into users’ personalities and routines. AI-driven behaviour prediction is a crucial component of surveillance capitalism. Its ability to translate raw data into behavioural data is precisely what makes AI technology so valuable in this context, because it allows companies to predict users’ behaviours with the highest degrees of accuracy. The ensuing practices are often justified on the grounds of more tailored advertising, richer service

offerings, or the free enjoyment of certain services. In this context, “Privacy is now less a line in the sand beyond which transgression is not permitted, than a shifting space of negotiation where privacy is traded for products, better services or special deals.”¹³

The resulting data commodification paradigm has been criticized for leaving individuals with no meaningful ways to consent to data collection, lack of legal protection regarding the inferences made from the bulk data collected, and lack of information regarding the processing and the parties involved.¹⁴ At a more abstract level, corporate surveillance has been criticized for taking away users’ capacity for judgement.¹⁵ The level of conditioning achieved through extreme content personalisation results in users gradually losing the ability to ponder over choice. In the long term, these mechanisms are detrimental to privacy and individual autonomy.¹⁶

¹⁰ *Evgeny Morozov*, The Real Privacy Problem, MIT Technology Review of 22 October 2013, available at: <https://www.technologyreview.com/2013/10/22/112778/the-real-privacy-problem/> (last visited 11 November 2025). See also, generally, *Jerome Joseph*, Big-data: catalyst for a privacy conversation, in: *Indiana Law Review* 48 (2014), pp. 213–242 (234). See also *Jones* (fn. 9).

¹¹ See, generally, *Yahya Alshamy et al.*, Surveillance Capitalism & the Surveillance State: A Comparative Institutional Analysis, in: *Constitutional Political Economy* 23 (2024), pp. 1–38.

¹² *Heather Suzanne Woods*, Asking more of Siri and Alexa: feminine persona in service of surveillance capitalism, in: *Critical Studies in Media Communication* 35 (2018), pp. 1–16 (12). See also *Hao-Ping Lee et al.*, Deepfakes, Phrenology, Surveillance, and More! A Taxonomy of AI Privacy Risks, CHI’24: Proceedings of the CHI Conference on Human Factors in Computing Systems, 11 May 2024, pp. 1–19 (10).

¹³ *Kevin D Haggerty/Richard Ericson*, The surveillant assemblage, in: *British Journal of Sociology* 51 (2000), pp. 605–622 (616).

¹⁴ See, generally, *Jane Andrew/Max Baker*, The General Data Protection Regulation in the Age of Surveillance Capitalism, in: *Journal of Business Ethics* 168 (2019), pp. 565–578.

¹⁵ *Laidler* (fn. 9). See also *Joseph* (fn. 10), p. 221.

¹⁶ It will be shown later that the privacy harms resulting from these mechanisms relate to decisional privacy and informational privacy. A definition of both of these values is proposed in the next section. On this point, see *Joseph* (fn. 10). See also, generally, *Lena Vatne Bjørlo*, Freedom from interference: Decisional privacy as a dimension of consumer privacy online, in: *AMS Review* 14 (2024), pp. 12–36. And see generally, *Yuxi Wu et al.*, The Slow Violence of Surveillance Capitalism: How Online Behavioral Advertising Harms People, FAccT ’23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (2023), pp. 1826–1837. And see *Daniel J. Solove*, Artificial intelligence and Privacy, in: *Florida Law Review* 77 (2025), pp. 1–73 (46).

Of course, the convergence of big data and AI technology raised concerns before the emergence of generative AI.¹⁷ AI technology was at work in the mechanisms involved in corporate surveillance very early on, in the form of predictive technology embedded into home and on-device assistants to gather data, and profiling algorithms designed to provide actionable insight into the habits of an individual and thus enable decision-making.¹⁸ Today, technological advances in the AI domain enable marketers and data scientists to collect more information, to make sense of larger volumes of data, and to infer granular knowledge about users.¹⁹ When it comes to generative AI in particular, the technology is notably used to power virtual companions²⁰ or digital versions of deceased loved ones.²¹ These applications are controversial for many reasons, including from a privacy standpoint, as they

may lead to the divulgence of very intimate data, thereby enabling higher levels of surveillance.

The wealth of data detained by the largest digital platforms makes them useful partners for governments. While the private sector may not always be aware of a state's tapping their databases, the private sector sometimes willingly cooperates with public agencies in state surveillance, repurposing commercial databases to accommodate the security needs of governments. For instance, China's state surveillance apparatus relies heavily on the private sector for the constitution of databases.²² In the famous NSA surveillance case, US Telecom company AT&T reportedly copied and transmitted the communications of its consumers to government authorities.²³ Similarly, in the facts leading to ECJ's "BCD case", bulk communications data (BCD) was collected by the Security and Intelligence Agencies from mobile network operator,²⁴ and US Supreme Court's *United States v. Miller* case featured the communication of a bank's client information to US government agencies.²⁵

2. AI in state surveillance

The term "surveillance state" is used to describe a model of governance relying on

¹⁷ Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes, 2019, available at: <https://search.coe.int/cm/?i=090000168092dd4b> (last visited 16 October 2025).

¹⁸ Laidler (fn. 9), p. 6; For a definition of 'profiling' see Art. 4 para. 4 of Regulation (EU) 2016/679 of 4 May 2016, OJ L 119, p. 1. For a more in-depth account on profiling, see Klaus Wiedemann, Profiling and (automated) decision-making under the GDPR: A two-step approach, in: Computer Law & Security Review 45 (2022), pp. 1-17 (3).

¹⁹ See Mireille Hildebrandt/Bert-Jaap Koops, The challenges of ambient law and legal protection in the profiling era, in: Modern Law Review 73 (2010), pp. 428-460 (435).

²⁰ Jessica Lucas, The teens making friends with AI chatbots, The Verge of 4 May 2024, available at: <https://www.theverge.com/2024/5/4/24144763/ai-chatbot-friends-character-teens> (last visited 6 November 2025).

²¹ Zeyi Yang, Deepfakes of your dead loved ones are a booming Chinese business, MIT Technology Review of 7 May 2024, available at: <https://www.technologyreview.com/2024/05/07/1092116/deepfakes-dead-chinese-business-grief/> (last visited 6 November 2025).

²² See, generally, Fan Liang et al., Constructing a Data-Driven Society: China's Social Credit System as a State Surveillance Infrastructure, in: Policy & Internet Special Issue: Social Media and Big Data in China 10 (2018), pp. 415-453.

²³ NSA Spying, Electronic Frontier Foundation, available at: <https://www EFF.org/fr/nsa-spying> (last visited 6 November 2025).

²⁴ ECJ, judgement of 6 October 2020, Case C-623/17, para. 25.

²⁵ Supreme Court of the United States of America, *United States v. Miller*, judgement of 21 April 1976, 425 U.S. 435 (1976).

pervasive surveillance tools to collect and analyse information about citizens for the purpose of anticipating crime, securing public spaces, and, more broadly, maintaining national security. State surveillance may be conducted through either digital or analogue means, although the increase in international terrorism in the 2000s and the subsequent digitization of society led to a generalization of the recourse to digital surveillance techniques. A prominent illustration of the surveillance state model is the extensively-documented national surveillance apparatus of the National Security Agency (NSA).²⁶

Corporate surveillance and state surveillance share some similarities, the first being the mechanisms at play *i.e.* the massive collection and analysis of data, usually implicating AI solutions. Second, the imbalance of power that characterizes the relationship between individuals, states and private actors means that the former are usually left with few means to resist surveillance, let alone the coercive power it enables. Third, the two surveillances may have a negative impact not only on privacy, but also on other fundamental rights such as free speech. Finally, there are no *pure* surveillance capitalists nor *pure* surveillance states, but a handful of business and governance models involving varying degrees of privacy intrusion.

The end goals are however dissimilar between the two types of surveillance. Indeed, while corporate surveillance seeks economic advantage by steering positive behaviour, state surveillance seeks national stability by discouraging them. Of the two, state surveillance may appear

more justifiable, which has notably led some authors warning against thinking of surveillance as a “malign plot hatched by evil powers.”²⁷ For instance, state surveillance was useful in the context of the spread of the coronavirus during the pandemic. Yet, as pointed by Solove, “Too much social control, however, can adversely impact freedom, creativity, and self-development.”²⁸ In the same vein, the independent high-level expert group on artificial intelligence appointed by the EU Commission emphasized the delicate process of striking a balance between the prevention of harm through surveillance practices and the protection of privacy and autonomy.²⁹

Much like the former, state surveillance is the subject of increasing attention because of the growing reliance of states on AI surveillance tools.³⁰ AI technology is at play in several mechanisms of state surveillance, where it can be used to perform various image processing tasks such as object and behaviour detection in order to predict scenarios, so-called “algorithmic surveillance.” Arguably more problematic, AI technology can also be leveraged to execute surveillance activities such as biometric identification, emotion recognition and biometric categorization for law enforcement.

The use of AI tools for the purpose of conducting state surveillance activities is problematic for a number of reasons. In 2023, the UN High-Level Advisory Body on Artificial Intelligence singled-

²⁶ David Lyon, Surveillance, Snowden, and Big Data: Capacities, consequences, critique, in: *Big Data & Society* 1 (2014), pp. 1-13 (2).

²⁷ Kirstie Ball *et al.*, *A Report on the Surveillance Society*, 2006, p. 4.

²⁸ Solove (fn. 6), p. 494.

²⁹ European Commission, *Ethics Guidelines for Trustworthy AI*, 8 April 2019, p. 13.

³⁰ Feldstein (fn. 8).

out real-time biometric surveillance for law enforcement purpose as posing an “unacceptable risk, violating the right to privacy.”³¹ Alongside the serious risk of biased outputs,³² one main issue with the inclusion of AI technology into a state’s surveillance apparatus is its ability to infer large quantities of information about physical persons based on the captured images. The French data protection authority speaks of a trend towards generalized “analysis”, as opposed to the initial generalized surveillance.³³ Such analysis leads to what Lyon calls “anticipatory governance”, where surveillance is “less concerned with the overall picture of a given individual as with ‘premeditating and pinpointing potential dangers.’”³⁴ In this context, the likelihood of errors and misuse is significant.

Unfortunately, the level of data transparency exhibited by surveillance activities is often lacking, preventing many from fully grasping the true extent of personal information private companies and states can extract from a few data points, how their data weighs in surveillance outcomes, and more broadly, the impact surveillance activities may have on their private lives. In this context, privacy

and by extension human autonomy and dignity, cannot be properly guaranteed.³⁵

III. Beyond data protection; making sense of privacy in the digital era

The right to privacy is considered “one of the foundations of a democratic society.”³⁶ This part offers a brief background to privacy, (1) before exploring the subset concept of digital privacy, (2) in order to explicate the nature of the legal harm resulting from surveillance activities.

1. Background to privacy

Given its prominent role in modern societies, the right to privacy is enshrined in many authoritative sources. At the international level, the right to privacy is enshrined in Art. 12 of the 1948 Universal Declaration of Human Rights³⁷ and Art. 17 of the 1966 International Covenant on Civil and Political Rights³⁸, both providing in identical terms; “No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks.”³⁹ The right to privacy

³¹ United Nations Advisory Body on Artificial Intelligence, Interim Report: Governing AI for Humanity, December 2023, para. 29.

³² *Jacob Snow*, Amazon’s Face Recognition Falsely Matched 28 Members of Congress with Mugshots, ACLU of 26 July 2018, available at <https://www.aclu.org/news/privacy-technology/amazons-face-recognition-falsely-matched-28> (last visited 20 October 2025).

³³ *Commission Nationale de l’Informatique et des Libertés (CNIL)*, Position sur les conditions de déploiement cameras dites “intelligentes” ou “augmentées” dans les espaces publics, 2022, p. 9.

³⁴ *Lyon* (fn. 26), quoting *Marieke de Goede*, The politics of privacy in the age of pre-emptive security, in: *International Political Sociology* 8 (2014), pp. 100–104 (102).

³⁵ *Bjørlo* (fn. 16).

³⁶ UN Doc. A/HRC/RES/28/16, p. 2.

³⁷ Universal Declaration of Human Rights of 10 December 1948, UN Doc. A/RES/217 A (III) (UDHR), Art.12.

³⁸ International Covenant on Civil and Political Rights of 16 December 1966, UNTS vol. 999, p. 171 (ICCPR), Art. 17.

³⁹ *Ibid.*; UDHR, Art. 12.

is also replicated in Art. 7 of the Charter of Fundamental Rights of the European Union⁴⁰, Art. 8 of the European Convention on Human Rights⁴¹, Art. 21 of the ASEAN Human Rights Declaration⁴², and Art. 11 of the American Declaration of the Rights and Duties of Man⁴³, among other important documents.

Despite its ubiquity, the right to privacy remains an elusive concept. In 2006 Solove observed “[P]rivacy is a concept in disarray. Nobody can articulate what it means”⁴⁴ Almost twenty years later, privacy remains “a complicated concept to review”⁴⁵ This is due to the fact that privacy is inherently a protean concept. Privacy applies both horizontally, in person-to-person settings, and vertically, in institutions-to-person settings. Each context brings different expectations towards the conduct of external parties.⁴⁶

In legal doctrine, privacy is apprehended simultaneously as a primary right susceptible of direct violation and as a source of more specific rights, the violation of which doubles as privacy infringement, such as with the right to protect reputa-

tion or the right to abortion.⁴⁷ The right to privacy is also an enabler of other rights and values. Hence, free speech, consumer protection and the right to public participation cannot exist without sufficient privacy guarantees.⁴⁸ But its intricate relationship with other rights and freedoms is not the sole source of difficulties when it comes to defining the concept of privacy.

Indeed, privacy and the protections stemming from it are in constant evolution. The social demand for privacy is itself in a state of flux, and varies based on societal, cultural and technological factors.⁴⁹ In addition, the right to privacy is not absolute because it admits exceptions that may differ from one domestic system to another. In fact, while, it would seem that most states are aware of the importance of safeguarding privacy, they do not necessarily approach it the same way. Take the example of the freedom of the press, which, until recently, was given primacy over privacy in the UK, while privacy had long prevailed over the freedom of the press in France.⁵⁰

⁴⁰ European Union, Charter of Fundamental Rights of the European Union of 14 December 2007, 2012/C 326/02, Art. 7.

⁴¹ Council of Europe, European Convention on Human Rights of 4 November 1950, as amended by Protocols Nos. 11, 14 and 15, ETS No. 005, Art. 8.

⁴² Association of Southeast Asian Nations (ASEAN), ASEAN Human Rights Declaration of 18 November 2012, Art. 21.

⁴³ Inter-American Commission on Human Rights (IACHR), American Declaration of the Rights and Duties of Man of 02 May 1948, Art. 11.

⁴⁴ Solove (fn. 6), p. 477.

⁴⁵ See, generally, Ali ALibeigi/Abu Bakar Munir/Md Ershadul Karim, Right to Privacy, a Complicated Concept to Review, in: Library Philosophy and Practice (e-journal) 2019, pp. 2841–2876.

⁴⁶ See notably Joseph (fn. 10), p. 234.

⁴⁷ See, among many other, ECtHR, *Pretty v. The United Kingdom* (2346/02), judgement of 29 April 2002, para. 61; US Supreme Court, *Roe v. Wade*, 410 U.S. 113, decision of 22 January 1973, para. 79.

⁴⁸ See notably UNHRC, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression of 17 April 2013, UN Doc. A/HRC/23/40, para. 24.

⁴⁹ Lee A. Bygrave, Privacy and Data Protection in an International Perspective, in: *Scandinavian Studies in Law* 56 (2010), pp. 166–200 (174). Samuel D. Warren/Louis D. Brandeis, *The Right to Privacy*, in: *Harvard Law Review* 4 (1890), pp. 193–220 (195). See also *ibid.*

⁵⁰ Kathryn F. Deringer, Privacy and the Press: The Convergence of British and French Law in Accordance with the European Convention of Human Rights, in: *Penn State International Law Review* 22 (2003), pp. 191–211 (192).

Few legal instruments provide a clear definition of privacy. As remarked by the European Court of Human Rights (ECtHR) in *Pretty v. The U.K.*; “[...] the concept of ‘private life’ is a broad term not susceptible to exhaustive definition.”⁵¹ Solove similarly states that “the term ‘privacy’ is an umbrella term, referring to a wide and disparate group of related things.”⁵² Be that as it may, the basic premises of privacy remain relatively discernible.

2. Digital privacy defined

Overall, privacy cases around the world have drawn from three theories of privacy; non-intrusion, self-determination, (or non-interference), and control over one’s information.⁵³ Each theory highlights one dimension of privacy: the *physical*, the *decisional* and the *informational* dimension, which complement each other in different ways.⁵⁴ Exploring pivotal case law on privacy provides valuable insights into its three dimensions.

In 1890 US lawyers Samuel D. Warren and Louis Brandeis famously referred to the right to privacy as the “right to be let alone”,⁵⁵ a rather rudimentary understanding of privacy mainly interpreted in the

context of the relationship between the administration and individuals. The concept was directly drawn from the fourth US constitutional amendment, which recognizes the right “[...]to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, [...]”⁵⁶ The view was later criticized for being “both too broad and too narrow to count as a successful definition.”⁵⁷ Still, the definition evokes the physical dimension of privacy: a safeguard for the physical premises of a person, or *physical privacy*.

Privacy does not only relate to the physical premises of a person, but extends to intangible values, such as one’s ability to make choices and take decisions regarding intimate matters without interference, or *decisional privacy*.⁵⁸ Building on a long-standing jurisprudence, the US Supreme Court for example considered in the oft-cited 1973 *Roe v. Wade* case that the right to abortion was encompassed in the right to privacy.⁵⁹ Similarly, decisional privacy has been an important part of the jurisprudence of the European Court of Human Rights stemming from Art. 8 ECHR, notably in connection to family matters.⁶⁰ At present, it is receiving increasing inten-

⁵¹ ECtHR (fn. 47), para. 61. And see ECtHR, *Niemietz v. Germany* (13710/88), judgement of 16 December 1992, para. 29. See also, an analysis of the relevant jurisprudence, at *Raphaël Gellert/Serge Gutwirth*, The legal construction of privacy and data protection, in: *Computer Law & Security Review* 29 (2013), pp. 522–530. For a thorough analysis of the different theories of privacy, see *Herman T. Tavani*, Philosophical Theories of privacy: implications for an adequate online privacy policy, in: *Metaphilosophy* 38 (2007), pp. 1–22 (6).

⁵² *Solove* (fn. 6), p. 485.

⁵³ *Tavani* (fn. 51), p. 7.

⁵⁴ *Bjørlo* (fn. 16).

⁵⁵ *Warren/Brandeis* (fn. 49), p. 193.

⁵⁶ Constitution of the United States of America of 17 September 1787, Amendment IV.

⁵⁷ *James H. Moor*, The ethics of privacy protection, in: *Library Trends* 39 (1991), pp. 69–82 (71). See also *Tavani* (fn. 51).

⁵⁸ *Ibid.*, p. 72. See also *Tavani* (fn. 51), p. 6.

⁵⁹ US Supreme Court, *Roe v. Wade*, 410 U.S. 113, decision of 22 January 1973.

⁶⁰ See for instance, ECtHR, *Schalk and Kopf v. Austria* (30141/04), judgement of 24 June 2010. See also *Bart van der Sloot*, Decisional privacy 2.0: the procedural requirements implicit in Art. 8 ECHR and its potential impact on profiling, in: *International Data Privacy Law* 7 (2017), pp. 190–201.

tion in the context of corporate surveillance.⁶¹

The third dimension of privacy, *informational privacy*, applies to situations implicating personal information, for instance, when one's reputation is damaged by a smear campaign or when one's personal data are being harvested without consent. Informational privacy is defined as the ability to exercise control over one's personal information, including image, correspondence and personal data. Informational privacy is sometimes referred to as "informational self-determination" or "informational autonomy."⁶² The informational dimension of privacy was, for example, in question in the 2017 *Bărbulescu v. Romania* case before the ECtHR, where the Court considered that instant message communications qualified as "correspondence" protected under Art. 8 ECHR, even when sent from the workplace. In *Whalen v. Roe*, the US Supreme Court makes explicit reference to a constitutional right to informational privacy.⁶³ Guarantees such as the protection of reputation and data protection principles including consent, control over the data, right to erasure, and rectification of information, relate to the informational dimension privacy. These principles are covered in most data protection laws.

In light of the above, it can be asserted that the right to privacy is a claim that extends to physical locations, the body, personal decisions, and digital as well as

non-digital information, as long as these closely relate to elements of the personality or the life of the rights-holder.⁶⁴

Digital privacy, in particular, both relates to the second and third aspects of privacy. Specifically, it motivates expectations regarding the ways third parties should treat the digital components of the private sphere, and what they effectively do with them, as long as the end goals have repercussions on their autonomy, taking into account the invasive nature of the technologies at play. In that sense, digital privacy acts as a defence against attempts to encroach on individual autonomy involving the processing of private information.

It bears noting that not all privacy invasions are blatantly illegal. In modern digital societies, individuals are invited to surrender components of their private selves in an ongoing manner, to access services, use goods, and overall, to improve their quality of life. Nevertheless, to relinquish personal information should not be equated to a total abandonment of privacy. It is helpful to consider the current paradigm as a sort of constant bargaining state, where components of the private self are exchanged for things *via* digital platforms. While absolutist views of informational privacy are hardly tenable under the current paradigm⁶⁵, privacy still requires that guarantees pertaining to the

⁶¹ *Bjørlo* (fn. 16).

⁶² BVerfGE 65, 1, 68–69.

⁶³ US Supreme Court, *Whalen v. Roe*, 429 U.S. 589, decision of 22 February 1977. For an analysis of the US case law dealing with informational privacy, see *Carlek Shachar/Carleen Zubrzycki*, Informational privacy after Dobbs, in: *Alabama Law Review* 75 (2023), pp. 1–50.

⁶⁴ See, for instance, ECtHR, *Perry v. the UK* (63737/00), judgement of 17 July 2003, para. 47. See also generally *Joseph* (fn. 10); US Supreme Court, *United States v. Jones*, 565 U.S. 400, decision of 23 January 2012; Concurring opinion of judge Sotomayor in *United States v. Jones*, 565 U.S. 400, decision of 23 January 2012; and see *Brandon T. Crowther*, (Un)Reasonable Expectation of Digital Privacy, in: *BYU Law Review* 2012, pp. 343–370.

⁶⁵ *Ibid.*, p. 237. See, for instance, *Florent Thouvenin*, Informational Self-Determination: A Convincing Rationale for Data Protection Law?, in: *Journal*

security and confidentiality of the information thus collected, and *in fine* the degree of autonomy kept by the right-holder, are assured.⁶⁶ In other words, the bargain should be based on trust and therefore, conditional.⁶⁷ Yet, as shown earlier, the privacy-adverse mechanisms involved in modern surveillance activities hardly, if ever, satisfy these criteria.⁶⁸

When consent is not required, as is often the case with state surveillance, the validity of surveillance practices must be assessed from the perspective of individuals' expectations of privacy, which must be reasonable. That is to say, balanced with the objectives sought and the necessity of the practice under scrutiny. Consent-based data collection practices should themselves guarantee free and informed consent. However, the practice leading to corporate surveillance usually rely on suboptimal measures to ensure informed consent. Few individuals realize the amount of personal information collected by corporations, let alone what is inferred from these data, and which decisions are taken based on said inferences. Due to their opacity and irresistibility, corporate surveillance practices subvert the conditions of the bargain, thereby undermining the concept of consent.⁶⁹

Besides, the commodification of attention and data that pervades surveillance activities inherently contradicts human dignity. As noted in the explanation paper to the Convention 108+, “[h]uman dignity requires safeguards to be put in place

when processing personal data, in order for individuals not to be treated as mere objects.”⁷⁰

IV. AI-driven surveillance in emerging AI laws

Are the mechanisms involved in surveillance – namely, the extensive accumulation and analysis of data, and the subsequent inferences drawn therefrom – adequately addressed in emerging legal regimes on AI? Leaving aside non-binding instruments, this section covers supranational regulatory initiatives pertaining to AI, identifying gaps in their approach to privacy (1), before expanding the discussion to other relevant regimes (2).

1. AI-driven surveillance in emerging international AI regulations

a) The AI Act

While considering relevant regulatory trends tailored to AI, one inevitably stumbles upon the recent EU's AI Act. Inspired by product safety rules, the AI Act sets forth a comprehensive and horizontal framework for the regulation of AI systems in the Union. Much like the General Data Protection Regulation (GDPR)⁷¹ before it, the AI Act could become a benchmark for AI regulation, the so-called “Brussel effect”.

of Intellectual Property, Information Technology and E-Commerce Law 2021, pp. 246-256.

⁶⁶ See notably *Solove* (fn. 6), p. 526.

⁶⁷ *Bjørlo* (fn. 16).

⁶⁸ *Bjørlo* (fn. 16).

⁶⁹ *Lyon* (fn. 34), p. 9.

⁷⁰ Additional Protocol to the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, regarding supervisory authorities and transborder data flows of 8 November 2001, ETS No. 181.

⁷¹ EU Regulation 2016/679 of 27 April 2017, OJ L 119 (GDPR).

The AI Act is based on the prescriptions set forth by the high-level expert group on artificial intelligence, which emphasized privacy as a core requirement to achieve “trustworthy AI”, understood as an AI system that is “lawful,” “ethical,” and “robust.”⁷² The text lays down several obligations for providers and deployers of AI systems, which vary based on the degree of risk associated with the system and its use. Some prohibitions are targeted at AI systems deemed to pose “unacceptable risks.” Importantly, the AI Act applies to both the public and the private sector, as long as the entity in question acts as provider or deployer of AI systems.

The AI Act does not deal specifically with data protection, since the GDPR already covers this important aspect of digital privacy. However, several privacy-adverse practices are addressed. Art. 5 AI Act notably prohibits particularly intrusive systems, which could be used for social control, and yield disproportionate harm to human rights, such as certain social-scoring practices and AI systems that create or expand “facial recognition databases through untargeted scraping of facial images.”⁷³

A first prohibition that seems relevant to corporate surveillance and its mechanisms relates to the use of AI systems for manipulative and exploitative purposes. The AI Act indeed prohibits some AI-enabled manipulative and exploitative practices involving the voluntary distortion of behaviours in ways that would

cause significant harm to a person or a group of persons.⁷⁴ Recital 29 of the AI Act, which provides some interpretative guidance on the matter, clarifies that the prohibition applies to the commercial context as well, but also notes that “common and legitimate commercial practices, for example in the field of advertising, that comply with the applicable law should not, in themselves, be regarded as constituting harmful manipulative AI-enabled practices.”⁷⁵

Beyond the unclarity introduced by the use of the adjectives “common and legitimate”, and although Recital 29 does not seem to evacuate completely the possibility that consumer manipulation practices akin to corporate surveillance fall into the scope of Art. 5 of the AI Act, the applicability of the relevant provisions is somewhat neutralised by a requirement of significant harm, or the likelihood thereof, which is hard to prove in the case of invasive advertising practices. How to measure the harm in the context of manipulative commercial practices remains unclear, although the Recital indicates that “unfair commercial practices leading to economic or financial harms to consumers are prohibited under all circumstances, irrespective of whether they are put in place through AI systems or otherwise.”⁷⁶

On this point, relevant provisions may also be found outside the AI Act, notably

⁷² European Commission (fn. 29), p. 2.

⁷³ AI Act, Recital 43. See also European Commission, Approval of the content of the draft Communication from the Commission - Commission Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act), C(2025) 884 final of 4 February 2025, p. 77.

⁷⁴ “The placing on the market, the putting into service or the use of certain AI systems with the objective to or the effect of materially distorting human behaviour, whereby significant harms, in particular having sufficiently important adverse impacts on physical, psychological health or financial interests are likely to occur, are particularly dangerous and should therefore be prohibited.”, AI Act, Recital 29, Art. 5 para. 1 lit. (a)(b).

⁷⁵ AI Act, Recital 29.

⁷⁶ *Ibid.*

in Directive 2005/29/EC (UCPD),⁷⁷ which, in its modernized form, notably protects European consumers against “coercion and undue influence”, understood as the act of “exploiting a position of power in relation to the consumer so as to apply pressure, even without using or threatening to use physical force, in a way which significantly limits the consumer’s ability to make an informed decision.”⁷⁸ Therefore, depending on their characteristics, targeted advertising practices could sometimes amount to undue influence, although this should also be appreciated in light of consumers’ own responsibilities.⁷⁹

The 2021 Guidance submitted by the European Commission on the interpretation and application of the UCPD offers a more in-depth analysis of the mechanisms involved in corporate surveillance. Data-driven practices, dark patterns and commercial practices of social media are notably addressed.⁸⁰ Interestingly, the guidelines acknowledge that the superior knowledge extracted at the data aggregation phase, the constant fine-tuning of commercial practices on consumers to learn more about their behaviour, as well as the opacity of the practices, may help to distinguish “highly persuasive advertising or sales techniques from, on the

other hand, commercial practices that may be manipulative and, hence, unfair under consumer law.”⁸¹ As the “significant harm” requirement of the AI Act is not replicated in the UCPD, it could be that its rules are easier to trigger than the AI Act’s; although it is likely that the threshold for recognising undue influence in the commercial context remains high outside of clear instances of coercion, as over-inclusive criteria risk outlawing the majority of business practices.

Early 2025, in conjunction with the priority entry into force of the provisions on prohibited practices, the European Commission published its Guidelines on prohibited artificial intelligence (AI) practices, as defined by the AI Act.⁸² The document notably covers the question of the scope of the prohibition enshrined in Art. 5 para. 1, distinguishing between lawful persuasion, which “operates within the bounds of transparency and respect for individual autonomy”, and manipulation, which involves “covert techniques undermining autonomy, leading individuals to make decisions they might not have otherwise made if they were fully aware of the influences at play.”⁸³ The guidelines adds that “Both the AI Act and the UCPD aim to proactively prevent consumer harm from AI-driven business practices that are manipulative, misleading, or aggressive”⁸⁴ and clarifies that the AI Act’s requirements are broader in scope than those of the UCPD in the sense that its provisions are not restricted to consumers and

⁷⁷ EU Directive 2005/29/EC of 11 May 2005, OJ L 149. (UCPD)

⁷⁸ *Ibid.*, Art. 9. See also EU Directive 2019/2161 of 27 November 2019, OJ L 328, Art. 3, adding transparency requirements as regard the nature of commercial search result. See Art. 2 for definitions.

⁷⁹ See in that sense European Commission, Commission Notice – Guidance on the interpretation and application of Directive 2005/29/EC of the European Parliament and of the Council concerning unfair business-to-consumer commercial practices in the internal market, OJ C 526 of 29 December 2021, pp. 99 ff.

⁸⁰ *Ibid.*

⁸¹ *Ibid.*

⁸² European Commission, Communication from the Commission - Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act), C(2025) 5052 final of 29 July 2025.

⁸³ *Ibid.*, para. 128.

⁸⁴ *Ibid.*, para. 136.

commercial harm.⁸⁵ But it transpires from the guidelines that the threshold set by the AI Act's requirements remains high. Ultimately, assessments will be made on a case-by-case basis, taking into account an array of parameters including transparency, conformity with data protection law, the vulnerability of the target, and the objective and impact of a technic, but the significant harm requirement makes it so that insidious techniques deployed by corporations to keep customers engaged with their product mostly fall outside the scope of the regulation, AI or not.

The AI Act is arguably more informative when it comes to state surveillance. Indeed, AI-driven social-scoring practices, which may be integrated in a state's surveillance apparatus⁸⁶ and lead to discriminatory and unjust decisions being taken to restrict the right of a person, are prohibited under the AI Act. The regulation is also concerned with surveillance practices involving biometric data in the form of biometric categorisation, real-time and post-remote biometric identification in publicly accessible spaces. The first two are in principle prohibited, while the third falls into the lower category of high-risk systems (Art. 26 para. 10 AI Act). This means that they are in principle authorised so long as some safeguards are in place. The AI Act also addresses profiling in the context of law enforcement. AI systems can be involved in profiling and decision-making processes, significantly contributing to the outcome, with potentially high impact on fundamental rights.⁸⁷ Art. 5 AI Act therefore prohibits, with exceptions, risks assessments and

crime prediction when based solely on profiling.⁸⁸

Finally, additional privacy-related requirements are contained in the regime for high-risk AI systems. Art. 10 AI Act indeed contains provisions on data governance, which may have some relevance to surveillance activities. Notably, Art. 10 para. 5 AI Act provides for the possibility to process special categories of data in order to mitigate biases in the outputs of an AI system. In this case, the AI Act demands that adequate privacy enhancing measures are deployed to complicate reidentification. Overall, bias reduction measures contribute to avoid unjust surveillance outcomes that may impact decisional privacy.

Be that as it may, the AI Act has been criticised for its permissive posture on real-time and post remote biometric identification, and overall lack of operational guidance. Particularly, the fact that the text still allows real-time remote biometric identification in “exhaustively listed and narrowly defined situations, where the use is strictly necessary to achieve a substantial public interest, the importance of which outweighs the risks”⁸⁹ has been described by human rights advocates as providing a “‘blueprint’ for how to conduct biometric mass surveillance practices” rather than strong privacy safeguards.⁹⁰

⁸⁸ See also AI Act, Recital 42.

⁸⁹ AI Act, Recital 32.

⁹⁰ European Digital Rights, How to fight Biometric Mass Surveillance after the AI Act: A legal and practical guide, EDRI of 27 May 2024, available at: <https://edri.org/our-work/how-to-fight-biometric-mass-surveillance-after-the-ai-act-a-legal-and-practical-guide/> (last visited 7 November 2025). See also *Laura Lazaro Cabrera*, EU AI Act Brief – Pt. 2, Privacy & Surveillance, Center for Democracy and Technology (cdt) of 30 April 2024, available at:

⁸⁵ *Ibid.*, para. 136.

⁸⁶ *Liang et al.* (fn. 22).

⁸⁷ *Wiedemann* (fn. 18).

Others have argued that the more lenient stance on post-remote biometric identification could be easily abused by authorities.⁹¹ One might also regret the fact that AI uses in the context of national defence are excluded from the scope of the regulation, thereby leaving the door open to misclassification and *in fine* abuse.

b) The framework convention on AI

The Framework Convention, adopted in May 2024 by the Council of Europe and opened to signature in September 2024, is the second most influential development in the field of international AI regulation to date. The text, which constitutes the first binding convention on AI with international reach, is intended to apply to “the activities within the lifecycle of artificial intelligence systems that have the potential to interfere with human rights, democracy and the rule of law” (Art. 3 para. 1 Framework Convention). On this point, the text appears to have a broader scope than the AI Act.

The Framework Convention is the result of the work of the Committee on Artificial Intelligence (CAI), based on preliminary research carried out by the Ad hoc Committee on Artificial Intelligence. Its drafting involved the 46 Member States of the Council as well as 11 observer States, including Japan and the United States, and 68 representatives of civil society. Like the AI Act, the Convention aims to promote human rights friendly AI, by adopting a risk limitation approach (Art. 1 lit. b). However, the Convention does not create new rights but sets out a number of general principles such as human dignity and personal autonomy (Art. 7), transparency

and control (Art. 8) and equality and non-discrimination (Art. 10), which draw directly from the guiding principles issued by the OECD.

As a *framework* Convention, the text seeks first and foremost to lay the foundations for more far-reaching international regulations in the future. Consequently, the Framework Convention is less technically detailed than the majority of national and European frameworks on the subject. The drafters chose not to name any specific activity involving AI that would fall within the scope of the text, leaving considerable room for manoeuvre for States to achieve its aims.

The Framework Convention on AI was also subject of criticism. Its broad formulation does not forecast strong effectiveness, which has led to it being heavily criticised, notably by the European Data Protection Committee (EDP).⁹² Yet, the general wording of its provisions is a consequence of its openness, as the CAI seeks to bring together States with different legal traditions, particularly in terms of AI regulation, which requires significant concessions. A follow-up mechanism provided in the form of a “Conference of the Parties” grants the Convention some degree of adaptability. However, it appears unlikely that a framework specifically addressing AI-driven surveillance activities will later be developed under the Convention. The

<https://cdt.org/insights/eu-ai-act-brief-pt-2-privacy-surveillance/> (last visited 7 November 2025).

⁹¹ Ibid.

⁹² EDPS statement in view of the 10th and last Plenary Meeting of the Committee on Artificial Intelligence (CAI) of the Council of Europe drafting the Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law of 11 March 2024, available at https://www.edps.europa.eu/press-publications/press-news/press-releases/2024/edps-statement-view-10th-and-last-plenary-meeting-committee-artificial-intelligence-cai-council-europe-drafting-framework-convention-artificial_en (last visited 20 October 2024).

exception expounded in Art. 3 para. 2 and 4 Framework Convention makes it that AI systems used for surveillance activities could easily fall outside its scope if they are shown to relate “to the protection of national interests.” The case of corporate surveillance is similarly uncertain under the Convention, as States are free to decide whether national private actors should be bound by the provisions of the Convention.

Hence, the Convention’s initial contribution in limiting surveillance practices is very tenuous, even though many aspects of surveillance contradict the basic rationale laid down in the explanatory document to the Convention; “[A]ctivities within the lifecycle of artificial intelligence systems should not lead to the dehumanization of individuals, undermine their agency or reduce them to mere data points [...]”⁹³

2. Guidance from non-AI specific regimes

AI systems rely on data to function. Their development and use therefore implicate data processing activities.⁹⁴ Therefore, a discussion on AI regulation mobilizes way more frameworks than technology-specific regimes. The AI Act also hints several times at the GDPR which provides data subjects with several rights that are directly relevant to this discussion.⁹⁵ The text notably recognises a right to object to

data processing including profiling when for the purpose of direct marketing, (Art. 21 GDPR) a right not to be subject to fully automated decision-making producing legal effect (Art. 22 GDPR), as well as a right to information and transparency regarding the logic involved, the significance and the envisaged consequences of automated decision-making for the data subject, (Art. 13 to 15 GDPR). As per Art. 23 GDPR, these rights can be restricted, among other, for national security reasons. Additionally, principles such as purpose limitation and data minimization (Art. 5 para. 1 lit. (b)(c) GDPR) also impose checks on surveillance practices.

At this point, the GDPR has been extensively discussed in the literature. Authors have underlined its inadequacy when it comes to AI-enabled surveillance. Andrew and Baker for instance argue that the GDPR’s complacency toward anonymisation and pseudonymisation “incentivize the use, collection, and trade of behavioural and other forms of de-identified data”, thereby enabling surveillance practices.⁹⁶ In the same vein, Zarsky argues that the provisions contained in Art. 22 GDPR on fully automated decision-making could be easily sidestepped by a data controller.⁹⁷ He adds that Big Data capabilities challenge the distinction between the different categories of data contained in the GDPR, with the most sensitive data extrapolatable from regular information.⁹⁸ More generally, prominent commentators have argued that the “individual control” model, on which most data protection legislations are built, is doomed, because it fails to account for the power imbalance

⁹³ Council of Europe, Explanatory Report to the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, CETS No. 225 of 5 September 2024, para. 53.

⁹⁴ See notably GDPR Recital 72.

⁹⁵ *Andrew/Baker* (fn. 14). And see AI Act, Art. 2 para. 7.

⁹⁶ *Andrew/Baker* (fn. 14).

⁹⁷ *Tal Zarsky, Incompatible: The GDPR in the age of big data*, in: *Seton Hall Law Review* 47 (2016), pp. 995-1020 (1016).

⁹⁸ *Ibid.*, p. 1017.

between companies, states and individuals.⁹⁹ Joseph A. Cannataci, former UN Special Rapporteur on the right to privacy also deplored the EU's lack of competence in the field of national security, which impedes proper oversight of surveillance policies.¹⁰⁰ Finally, recent discussions in the realm of generative AI regulation have highlighted the GDPR's poor performance in capturing the particularities of generative AI systems.¹⁰¹

Much like the GDPR, the Convention 108+ modernizing the Convention 108 for the Protection of Individuals with regard to Automatic Processing of Personal Data covers diverse aspects of AI-driven surveillance activities. Adopted in 2018, the modernization Protocol for the Convention 108 provides broad guidelines for the international protection of data worldwide which integrates provisions directly targeted at AI systems.¹⁰² Unlike the first version, the amended Convention 108+ is fully applicable to the national security domain. It also applies to both the public and the private sector, which makes it a more impactful instrument than the Framework Convention

when it comes to controlling corporate surveillance.

The Convention 108+ constitutes the only existing binding treaty on privacy and data protection in the digital context. The CoE's Committee of Minister has made multiple references to the Convention 108+, among others, at the occasion of a non-binding declaration on risks arising from surveillance technologies¹⁰³, and a recommendation dealing with automatic processing of personal data in the context of profiling.¹⁰⁴ When it comes to data processing in the context of national security, the Convention 108+ requires a test of proportionality and necessity. The Convention takes up several principles enshrined in the 2014 International Principles on the Application of Human Rights to Communications Surveillance,¹⁰⁵ a document drafted by privacy experts aiming

⁹⁹ *Daniel J. Solove/Woodrow Hartzog, Kafka in the Age of AI and the Futility of Privacy as Control*, in: *Boston University Law Review* 104 (2024), pp. 1021-1042 (1031).

¹⁰⁰ UNHRC, Report of the Special Rapporteur on the right to privacy of 16 October 2019, UN Doc A/HRC/40/63.

¹⁰¹ *Juliette Sénéchal*, Publication de l'avis de l'EDPB du 17 décembre 2024 sur le traitement des données personnelles dans le contexte des modèles d'IA : prémices d'une mutation profonde du RGPD ?, *Dalloz actualités* of 17 January 2025, available at: <https://www.dalloz-actualite.fr/flash/publication-de-l-avis-de-l-edpb-du-17-decembre-2024-sur-traitement-des-donnees-personnelles-da> (last visited 7 November 2025).

¹⁰² Committee of Ministers of the Council of Europe, *Modernised Convention for the Protection of Individuals with Regard to the Processing of Personal Data*, CM/Inf (2018)15-final of 18 May 2018.

¹⁰³ Committee of Ministers of the Council of Europe, *Declaration of the Committee of Ministers on Risks to Fundamental Rights stemming from Digital Tracking and other Surveillance Technologies*, Decl(11/06/2013) of 11 June 2013.

¹⁰⁴ Committee of Ministers of the Council of Europe of the Council of Europe, *The protection of individuals with regard to automatic processing of personal data in the context of profiling Recommendation*, CM/Rec(2010)13 of 23 November 2010. See also Committee of Ministers of the Council of Europe, *Recommendation Rec(2002)9 on the protection of personal data collected and processed for insurance purposes*, Rec(2002)9 of 18 September 2002; Committee of Ministers of the Council of Europe of the Council of Europe, *Recommendation Rec(97)18 concerning the protection of personal data collected and processed for statistical purposes*, Rec(97)18 of 30 September 1997.

¹⁰⁵ *Juan Carlos Lara/Valentina Hernández/Katitza Rodríguez*, *International Principles on the Application of Human Rights to Communications Surveillance and the Inter-American System for the Protection of Human Rights of August 2026*, available at: <https://necessaryandproportionate.org/files/iachr-en-august2016.pdf> (last visited 17 November 2025).

to provide state actors with precise guidelines regarding the conduct of surveillance activities.

So far, the Protocol modernizing Convention 108+ has been ratified by 33 States, the majority being European. The updated Convention will only enter into force once this number reaches 38. It is worth noting that the United States, which hosts the most powerful digital firms, was not a party to the original Convention 108+. Given the current priorities at the White House, it is unlikely that Convention 108+ will be ratified by the US government.

V. Conclusive remarks on advancing AI privacy discussions

Privacy should play a central role in the regulation of AI tools. Current legal development on the matter however show that this is not really the case and that non-AI-specific frameworks have weaknesses. It is therefore interesting to investigate whether human rights law, which by default applies to AI technology and its uses, is up to the task of filling the gaps left by more specific frameworks when it comes to mitigating AI-enabled privacy risks. After all, both the Framework convention and the Convention 108+ limit their exceptions on national security to the respect of international human rights law. Unfortunately, due to several theoretical and structural deficiencies, international human rights law might not provide sufficiently robust baseline protection to individuals whose privacy is infringed upon by AI-driven surveillance practices.

On an abstract level, the inherent fluidity of the concept of privacy makes it difficult

to operationalize in practice. The absence of a clear definition for privacy and its subjective dimensions necessitate an ongoing evaluation of the numerous expectations stemming from it. Admittedly, privacy must be considered in context, and approached as a mutable concept. It must be able to satisfactorily respond to new challenges and mitigate harms to human dignity and autonomy while simultaneously allowing society to function. But privacy cannot be toned down on the basis that individuals are giving up so many of it nowadays. Human autonomy and dignity are invariable, and as such, should always guide assessments of privacy expectations.

At present, it is difficult for individuals to understand when their data is used for AI-training purposes, especially since the issue is relatively new, large databases already exist and so is a sense of resignation over the propriety of personal data.¹⁰⁶ While the constant bargaining taking place online is a source of privacy risks that the principle cannot eliminate entirely, societies cannot afford to allow countervailing considerations to prevail over privacy in the constant checks and balances imposed by ubiquitous computing environment.¹⁰⁷ This means, first and foremost, that data collection must be rationalized, in the sense of empowering the data subject to make free decisions regarding the amount of personal information that is relinquished in exchange for a ser-

¹⁰⁶Nora A. Draper/Joseph Turow, The corporate cultivation of digital resignation, in: *New Media & Society* 21 (2019), pp. 1824-1839 (1831).

¹⁰⁷Stefan G. Weber/Andreas Heinemann/Max Mühlhäuser, Towards an Architecture for Balancing Privacy and Traceability in Ubiquitous Computing Environments, paper presented at Third International Conference on Availability, Reliability and Security, 4-7 March 2008, pp. 958-964.

vice or goods. In situations where consent can be circumvented, it is imperative to ensure transparency with regard to the collection, use and the anticipated outcomes of such data processing operations.

Still, a blatant issue with privacy as enshrined in the various existing international documents is that it is centred around the individual and thus, struggles to accommodate collective needs related to data processing. If anything, surveillance activities are societal-scale undertakings. Profiling virtually concerns billions of users. At the individual level, the right to privacy may offer some degree of protection against data misuses, but it cannot address the full picture. Indeed, due to the situation of quasi-monopoly of a few companies, users do not really possess a negotiating power regarding the trade of data for services. This power imbalance, which favours the acceptance of companies' terms and conditions, counteracts any claim of arbitrariness and, ultimately, limits the relevance of the right to privacy as a safeguard since data subjects more often than not enter into data transaction without a proper understanding of the implications.¹⁰⁸ Even when basic privacy requirements are satisfied, the content of the right to privacy becomes gradually shallower as more data is required to access common services, and more data-sensitive activities are integrated in everyday life interactions.

The very individualistic and consent-oriented understanding of privacy as enshrined in international instruments is therefore lacklustre. Further advance in the protection of users' data against corporate surveillance will not come from the current approach, but from rebalancing the bargain between users and

¹⁰⁸*Solove/Hartzog* (fn. 99).

providers. As expressed by some authors; "A privacy and data protection framework that places the primary responsibility on individuals to manage their data across hundreds, even thousands, of digital relationships and channels fundamentally does not scale, and thus will not succeed in protecting individual privacy."¹⁰⁹ The concern was recently echoed by Solove and Hartzog, who called for the application of a "societal structure" model of privacy regulation also embracing AI.¹¹⁰

Another noticeable impediment to the performance of a human rights framework in the present case relates to the fact that human rights law is principally state-centric. While states are directly bound by the human rights treaties to which they commit, in addition to certain customary human rights which are binding upon all states, private entities are not directly bound by international human rights law. This is why State participation in instruments such as the International Covenant on Civil and Political rights is crucial. Although modern developments in the realm of human rights law have recognized the human rights responsibilities of private actors,¹¹¹ these actors are only liable for human rights harm under national law. It is thus the primary responsibility of states to ensure that the right to privacy is respected within their borders.

¹⁰⁹*Jennifer King/Caroline Meinhardt*, Rethinking Privacy in the AI Era: Policy Provocations for a Data-Centric World, White Paper of 22 February 2024, available at: <https://hai.stanford.edu/policy/white-paper-rethinking-privacy-ai-era-policy-provocations-data-centric-world> (last visited 20 October 2025), p. 30.

¹¹⁰*Solove/Hartzog* (fn. 99).

¹¹¹United Nations Human Rights Office of the High Commissioner, Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework, HR/PUB/11/4 (Ruggie Principles) of 2011.

In addition, business firms usually enjoy greater freedom when it comes to data practices under the right to the freedom to conduct business, which may be used as a defence against certain claims. As a result, attempts to circumscribe the data practices of the private sector have been paradoxically weaker than for state surveillance.

Although regulators can influence the fairness of the bargain, their action is limited for several reasons. First, it might be difficult for regulators, to assess what is necessary when providers offer a wide range of services requiring various data to function properly, such as location and browsing data. Second, regulators might feel pressure to avoid undermining innovation and development in the digital sector, especially when national firms are concerned. Third, the promise of better population control through the use of AI technology is inherently attractive for authorities, and effectively curtails the right to privacy.

Nevertheless, the balance of interests between individuals, corporations and states must be readjusted, and new approaches to privacy might be the key. This difficult endeavour can only stem from national or regional initiatives, since value decisions are beyond the scope of international human rights law. To this end, the current momentum around AI regulation should be exploited fully. The longstanding paradigm revolving around the commodification of personal information for the purpose of influencing behaviours, especially when the objectives sought are of economic nature, needs to be challenged.

Vitae

Dr. William Letrone is a CNRS postdoctoral researcher in cybersecurity and data protection law at Nantes University, France and a member of the IPoP project.

Dr. Tony Cabus is a postdoctoral researcher at the Walther-Schücking Institute for International Law in Kiel, Germany.

Responsibility and Accountability for the Use of AI in Law Enforcement in the European Union – Lost in Negotiations?

Steven Kleemann¹  • Milan Tahraoui^{2,3} 

¹University of Potsdam, Law Faculty

²Centre Marc Bloch (Berlin)

³Paris 1 Pantheon-Sorbonne University

Contents

- I. Introduction
- II. The Area of Law Enforcement in a Risk-Based Approach
- III. Accountability Mechanisms under the AI Act and AI contestability
- IV. Conclusion
Vitae

Abstract

This paper examines the EU AI Act's application to law enforcement, highlighting how this sector is incorporated into the risk-based approach and assessing the extent to which such incorporation could weaken safeguards for individuals. It argues that, although the newly created accountability framework is complex, it offers only limited remedies for affected individuals. To ensure genuine protection of fundamental rights, the exceptions ('backdoors') embedded in the framework must be critically examined, contestability mechanisms must be strengthened, and the responsibilities of providers and deployers of high-risk AI must be clarified. Where appropriate, a rights-based approach should be integrated into the risk-based approach to underscore that fundamental rights are non-negotiable. This integration is essential to align the use of AI with the AI Act's twin objectives of protecting fundamental rights and promoting innovation.

Keywords

AI Act, Law Enforcement, Accountability, Contestability, Risk Regulation, Fundamental Rights, Digital Policy

Citation:

Steven Kleemann/Milan Tahraoui, Responsibility and Accountability for the Use of AI in Law Enforcement in the European Union, in: MRM 30 (2025) 2, pp. 117–143. <https://doi.org/10.60935/mrm2025.30.2.28>.

Received: 2025-08-23

Accepted: 2026-01-09

Published: 2026-02-17

Permissions:

The copyright remains with the authors.
Copyright year 2026.

Unless otherwise indicated, this work is licensed under a [Creative Commons License Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/). This does not apply to quoted content and works based on other permissions.

I. Introduction*

The European Union Regulation laying down harmonised rules on artificial intelligence (AI Act)¹ entered into force on 1 August 2024. Despite the formal adoption of the AI Act, this contribution will mainly anticipate its future application in practice, as the Act set forth differentiated dates for the entry into force of its various chapters, sections and provisions.

In December 2023, prior to its adoption, the European Parliament, the Council of the European Union and the European Commission reached a compromise during the trilogue negotiations.² In this context, the regulation of the use of AI for law enforcement activities was one of the most controversial issues in the negotiations around the AI Act, along with the issue of so-called foundation AI systems, now called general-purpose AI models, which culminated in a three-day marathon trilogue process.³ Although the dangers posed by AI systems are being addressed

more frequently, at least nominally in international policy and legal documents, the risks posed by the increasing use of AI tools for law enforcement purposes have often been narrowly focused on the – albeit undeniably important – aspects of data protection, the legal regulation of data processing and the regulation of facial recognition technologies. However, the use of AI systems by law enforcement agencies raises further questions, particularly concerning the risk of fundamental rights being violated by AI-based decisions or actions. This is especially true for the aspects of the AI Act that have come under strong criticism from the perspective of fundamental rights protection.

Since its final adoption, these criticisms have been somewhat vindicated, as the European Commission has committed itself to a so-called simplification process⁴ with the potential to further weaken this

* This paper is partially based on the research project “VIKING” (FKZ 13N16242), funded by the German Federal Ministry of Education and Research (BMBF), now the Federal Ministry of Research, Technology and Space (BMFTR). The authors wish to thank Mario Petoshati for his assistance with proofreading and editing this paper. This article is based on a paper presented at the “30th Anniversary Conference Human Rights and Artificial Intelligence Addressing challenges, enabling rights”, 7-8th November 2024 in Potsdam, Germany.

¹ EU Regulation 2024/1689 of 12 July 2024, OJ L, 2024/1689 (AI Act).

² For an overview on the notion of trilogue in the European Union, see *Giacomo Ruge*, *Trilogues: the democratic secret of European Legislation*, 2025.

³ For an account of the main controversies that took place during the negotiations of the AI Act, see *European Digital Rights*, *EU AI Act Trilogues: Status of Fundamental Rights Recommendations*, EDRI of 16 November 2023, available at:

<https://edri.org/our-work/eu-ai-act-trilogues-status-of-fundamental-rights-recommendations/> (last visited 9 December 2025); *Jeremy Fleming-Jones*, *EU AI Act nearing agreement despite three key roadblocks – co-rapporteur*, euronews of 23 October 2023, available at: <https://www.euronews.com/next/2023/10/23/eu-ai-act-nearing-agreement-despite-three-key-roadblocks-co-rapporteur> (last visited 15 September 2025); *Müge Fazlioglu*, *Contentious areas in the EU AI Act trilogues*, IAPP News of 30 August 2023, available at: <https://iapp.org/news/a/contentious-areas-in-the-eu-ai-act-trilogues> (last visited 15 September 2025).

⁴ See *European Commission*, *Simplification*, 2025, available at: https://commission.europa.eu/law/law-making-process/better-regulation/simplification-and-implementation/simplification_en (last visited 21 August 2025); *Sarah Chanter/Caterina Rodelli*, *One Year On, EU AI Act Collides with New Political Reality*, Tech Policy Press of 7 August 2025, available at: <https://www.techpolicy.press/one-year-on-eu-ai-act-collides-with-new-political-reality/> (last visited 21 August 2025).

protection.⁵ Many of the obligations contained in this regulation are either vaguely worded or, even if specific requirements are well-defined, they contain broad exemptions for law enforcement. Furthermore, the final version of the AI Act incorporates a somewhat limited perspective on contestability, in a broader sense, regarding the impact of AI systems on affected persons, due to the fact that it originally was primarily drafted on the basis of pre-existing EU product safety law.

Moreover, the European Commission and other European institutions are under considerable pressure to alleviate the alleged regulatory burden placed on firms and industry actors,⁶ who criticize the regula-

tory model of the AI Act for impeding innovation and preventing security threats from being addressed.⁷ The high-risk AI requirements in Articles 8–27 of the AI Act are central to safeguarding fundamental rights, especially where law enforcement authorities deploy AI systems. The obligations affect providers, distributors and deployers differently. Although Articles 8–15 AI Act do not always specify addressees, they largely concern providers given their focus on system design and development⁸—although this is not the case systematically.⁹ This allocation of re-

⁵ As this paper was written before the Digital Omnibus Proposal was officially announced by the European Commission, we do not address the many implications of the so-called simplification, which has sparked controversy, particularly with regard to the GDPR and the AI Act. See, European Commission, Digital Omnibus Regulation Proposal of 19 November 2025, available at: <https://digital-strategy.ec.europa.eu/en/library/digital-omnibus-regulation-proposal> (last visited 1 December 2025); European Commission, Digital Omnibus on AI Regulation Proposal of 19 November 2025, available at: <https://digital-strategy.ec.europa.eu/en/library/digital-omnibus-ai-regulation-proposal> (last visited 1 December 2025). However, the envisaged amendments by the European Commission could raise several contentious issues, including the weakening of responsibility and accountability frameworks regarding the use of AI for law enforcement purposes. See for instance, European Digital Rights, Press release: Commission’s Digital Omnibus is a major rollback of EU digital protections, EDRI of 19 November 2025, available at: <https://edri.org/our-work/commissions-digital-omnibus-is-a-major-rollback-of-eu-digital-protections/> (last visited 1 December 2025).

⁶ See for instance, European Digital Rights et al., Open Joint letter against the Delaying and Reopening of the AI Act, EDRI of 9 July 2025, available at: <https://edri.org/our-work/open-letter-european-commission-must-champion-the-ai-act-amidst-simplification-pressure/> (last visited 21 August 2025); European Center for Not-

for-Profit Law et al., Open Letter to the European Commission on the announced withdrawal of the AI liability, ECNL of 7 April 2025, available at: <https://ecnl.org/news/eu-needs-ai-liability-rules> (last visited 21 August 2025); *Melissa Heikkilä/Barbara Moens*, EU lawmakers warn against ‘dangerous’ moves to water down AI rules, Financial Times of 25 March 2025, available at: <https://www.ft.com/content/9051af42-ce3f-4de1-9e68-4e0c1d1de5b5> (last visited 24 October 2025); *Maria Maggiore/Leila Miñano/Harald Schumann*, France spearheads member State campaign to dilute Europe AI regulation, Investigate Europe of 22 January 2025, available at: <https://www.investigate-europe.eu/posts/france-spearheads-member-state-campaign-dilute-european-artificial-intelligence-regulation> (last visited 21 August 2025); *Francesca Palmiotta*, The AI Act Roller Coaster: The Evolution of Fundamental Rights Protection in the Legislative Process and the Future of Regulation, in: European Journal of Risk Regulation 16 (2025), pp. 770–793 (789 f.).

⁷ EU Champions AI Initiative, Stop the Clock – Open letter, July 2025, available at: <https://aichampions.eu/#stoptheclock> (last visited 21 August 2025).

⁸ Article 16 lit. a. AI Act states that providers must fulfil the high-risk obligations from Section 2 (Art. 8–15 AI Act), which suggests therefore that they are the main addressees. Art. 15–27 AI Act then differentiate between AI providers, AI distributors, AI importers, AI deployers and other parties involved.

⁹ See, for example, Art. 14 AI Act on human oversight, which mainly addresses AI providers but whose obligations can only be fulfilled in cooperation with AI deployers. This is among others indicated in Art. 26(3) AI Act, which deals with

sponsibilities can undermine fundamental rights protection: providers must establish a risk management system for high-risk AI (Art. 9), even though deployers, users or affected persons—particularly in policing contexts—may be better positioned to identify actual rights risks in practice.¹⁰ The AI Act partly compensates for this asymmetry by requiring deployers to conduct a fundamental rights impact assessment (Art. 27), which covers similar concerns, albeit through a different mechanism.

Regarding prohibited AI practices (Art. 5), both deployers and providers carry duties aimed at preventing unlawful interferences with rights. Deployers may neither use prohibited AI systems nor operate systems in ways that amount to a prohibited practice. Providers, in turn, must ensure that their systems cannot function—or be reasonably used—in prohibited ways. They must implement effective, verifiable and proportionate safeguards against foreseeable misuse, include contractual clauses banning unlawful applications, and provide clear guidance on correct use and the need for human oversight.¹¹ The

distribution of responsibilities reflects each actor’s control over design, development and deployment, and must be assessed proportionately for each prohibition to ensure that those best placed to prevent rights violating uses actually do so.¹²

By contrast, fewer analyses address the rights granted to individuals under the AI Act to contest AI-driven interferences—such as discriminatory policing tools—or to challenge AI development and deployment projects. Strengthening these avenues of redress will require further measures at the national level. Moreover, rights-based contestation interacts with institutional oversight mechanisms—supervisory authorities, complaint procedures, data protection processes and fundamental rights impact assessments. These mechanisms define obligations but must also be articulated in terms of their underlying protective function: the rights of affected persons, available remedies, and the practical requirements for independent and effective supervision. Many implementation challenges arise precisely at this intersection between institutional responsibilities and the need to secure enforceable fundamental rights protections against AI-based law enforcement practices.

Thus, this paper addresses the following research questions: how are particularly sensitive areas, such as the use of AI for coercive public security purposes like law enforcement, incorporated into the AI Act’s regulatory framework? What accountability and responsibility mechanisms are in place for the use of AI in these areas, with regard to the protection of fundamental rights for affected persons? To

the obligations of AI deployers and refers to “the deployer’s freedom to organise its own resources and activities for the purpose of implementing the human oversight measures indicated by the provider”. See about the distribution of roles between AI providers and AI deployers for the implementation of human oversight obligations, *Johan Laux/Hannah Ruschemeier*, Automation Bias in the AI Act: On the Legal Implications of Attempting to De-Bias Human Oversight of AI, in: *European Journal of Risk Regulation* 16 (2025), pp. 1519–1534 (1524 ff.).

¹⁰ *Nathalie A. Smuha/Karen Yeung*, The European Union’s AI Act: Beyond Motherhood and Apple Pie?, in: *Nathalie A. Smuha* (ed.), *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence*, 2025, pp. 228–258 (241 f.).

¹¹ European Commission, *Commission Guidelines on prohibited artificial intelligence practices es-*

established by Regulation (EU) 2024/1689 (AI Act), C(2025) 884 final of 29 July 2025, para. 40.

¹² *Ibid.*, para. 20.

what extent can these mechanisms sustain avenues of contestability against AI development and deployment in these sensitive areas? We will consider the advantages of formal and informal channels of contestability from a rights-based perspective. In this aim, Section II will elaborate on the area of law enforcement within a risk-based approach, including the definition of risk in the AI Act. Section III introduces some of the main features of responsibility and accountability mechanisms with a focus on contestability in its subsections, while Section IV concludes by providing an overview of the current gaps in the AI Act regarding meaningful fundamental rights protection for the use of AI in law enforcement, and briefly suggesting potential solutions.

II. The Area of Law Enforcement in a Risk-Based Approach

Policing, criminal justice, migration, asylum and border control management are not excluded from the scope of the AI Act, as opposed to the use of AI systems for military, defence and national security purposes. However, due to this integrated approach, special ‘backdoors’ are used to employ risky AI systems that would be prohibited or restricted if they were used by other state agencies or non-state actors.¹³

¹³ On the issue of ‘backdoors’ in greater detail, see: Steven Kleemann/Hartmut Aden, Die Nutzung Künstlicher Intelligenz durch Strafverfolgungsbehörden – „Hintertüren“ der Verordnung der Europäischen Union über Künstliche Intelligenz, in: Wilfried Honekamp/Stefanie Kemme/Jens Struck (ed.), Auswirkungen von KI auf die zukünftige Polizeiarbeit. Technologische Potenziale, rechtliche Rahmenbedingungen, kriminologisch-sozialwissenschaftliche Erkenntnisse, 2025, pp. 3–30.

Many of these ‘backdoors’ were introduced by the Council of the European Union in its amended version of the AI Act through interventions from representatives of the EU Member States as well as the security and law enforcement communities.¹⁴ The term ‘backdoor’ refers to a number of different approaches that could be taken to enable the use of risky AI systems. These include special exceptions from generally applicable requirements and enabling conditions for the use of these AI systems, or more indirect legal means. An example of this are the permissive rules contained in the AI Act for regulating the testing of AI systems in real-world conditions for law enforcement purposes (Art. 60 AI Act). Furthermore, an analysis of the AI Act clearly shows that there are even more ‘backdoors’ for the use of AI systems in the domains of migration, asylum and border management and control.¹⁵ Despite these different legal categorisations, the AI tools used in these two domains of law enforcement are indeed similar, with the major difference being that AI used for migration, asylum and border control management is more permissively regulated for national authorities than their use for law enforcement purposes.

¹⁴ Palmiotto (fn. 6), p. 780, p. 787, pp. 789 ff.; *Ludvine Sarah Stewart*, The regulation of AI-based migration technologies under the EU AI Act: (Still) operating in shadows?, in: *European Law Journal* 30 (2024), pp. 122–135 (127 f.).

¹⁵ This can be observed by comparing Annex III para. 6 f. of the AI Act. See also *Alberto Rinaldi/Sue Anne Teo*, The Use of Artificial Intelligence Technologies in Border and Migration Control and the Subtle Erosion of Human Rights, in: *International and Comparative Law Quarterly* 74 (2025), pp. 61–89 (83 f.), arguing that the lines between security and migration have been and continue to be increasingly blurred and that the AI Act “ended up compressing distinct State obligations relating to borders and migration into the same risk bucket”.

It can be argued that the inclusion of security agencies within the scope of the AI Act represents an attempt to overcome the opt-out logic that has characterised other EU policy areas, such as data protection. It is also noteworthy that accountability mechanisms in the form of remedies are now included in a separate Section 4 of Chapter IX of the AI Act (“Post-market monitoring, information sharing, and market surveillance”), despite no direct remedial mechanisms for persons affected by AI being foreseen in the European Commission’s original proposal. In addition, other obligations, such as the obligation to conduct a fundamental rights impact assessment for high-risk AI systems aim to contribute to accountability (Art. 27 AI Act).¹⁶

1. Risks as defined in the AI Act

The AI Act defines the term ‘risk’ in Art. 3 para. 2 as “the combination of the probability of an occurrence of harm and the severity of that harm”. Furthermore, the Act distinguishes between different levels of risk intensity. The AI Act essentially differentiates between four risk categories (unacceptable, high, limited and minimal or no risk), which impose varying requirements on such systems. Moreover, during the negotiations of the AI Act, the co-legislators introduced a new category of general-purpose AI models¹⁷ and the no-

tion of ‘systemic risk’¹⁸ for those AI models that can be qualified as such under Art. 51. However, this new category is at odds with a truly risk-based approach, and these ambiguities should be taken into account when further examining this topic.

It is first necessary to differentiate between risks and uncertainties. A risk may be defined by Art. 9 AI Act as a ‘known known’, containing statistical probabilities and quantifiable effects, which is reflected in the general definition of ‘risk’ under Art. 3 para. 2 AI Act: “the combination of the probability of an occurrence of harm and the severity of that harm”. This definitional take can be criticised “as the infringement of a fundamental right does not necessarily require any ‘harm’ to ensue”.¹⁹ An uncertainty is, in comparison, a ‘known unknown’, that is, a situation or event that cannot be quantified because the effects of a specific technology are not yet known. Furthermore, there are instances of ‘unknown unknowns’, whereby there is no awareness that certain things or activities may have negative effects, despite the potential for such effects to

the draft Communication from the Commission – Guidelines on the scope of obligations for general-purpose AI models established by Regulation (EU) 2024/1689 (AI Act), C(2025) 5045 final of 18 July 2025, paras. 12 ff.

¹⁶ Steven Kleemann/Hartmut Aden, Die Grundrechte-Folgenabschätzung nach dem Artificial Intelligence Act der Europäischen Union – eine Chance für die polizeiliche KI-Nutzung, in: Sabrina Schönrock/Hartmut Aden (ed.) Breitscheidplatz-Symposium 2024: Zukunftssicherheit: Die Rolle von KI im Kampf gegen den Terrorismus, pp. 47-57 (48 ff.).

¹⁷ See AI Act, Chapter V, Arts. 50-56, pp. 83-87; European Commission, Annex to the Communication to the Commission, Approval of the content of

¹⁸ Art. 3 para. 65 AI Act defines it as: “a risk that is specific to the high-impact capabilities of general-purpose AI models, having a significant impact on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain.”

¹⁹ Nathalie A. Smuha, The paramountcy of data protection law in the age of AI (Acts), in: European Data Protection Supervisor (ed.), Two decades of personal data protection. What next? – EDPS 20th Anniversary, 2024, pp. 225-239 (235).

exist.²⁰ In this regard, it can be argued that the deployment of AI systems for border controls is based on a surveillance logic aiming at discovering ‘unknown unknown’ security risks, by inferring information or creating *sui generis* risk group profiles based on AI technologies.²¹ Consequently, it is crucial to achieve consensus on the selection of risks and the severity assigned to them, whether in terms of probability, impact, or both, in order to ensure the success of any risk-based approach.²² This discussion was particularly pertinent during the genesis of the regulation, when there was an intensive debate about whether specific AI systems or applications should be banned in the EU. There were also fundamental debates regarding the classification of some AI technologies as high-risk, minimal risk, or as the relatively new classification of general-purpose AI models.

Those debates surrounding classifications were of particular importance, as they imply different regulatory consequences. For example, AI systems that are classified as high-risk use-cases, must comply with essential, specific, and procedural precautions. This builds upon various issue areas in which EU law attempts to regulate risks.²³ For the regulation of high-risk ac-

tivities, the EU has adopted the so-called precautionary principle to regulate risks, notably in environmental policy²⁴ (Art. 191 para. 2 TFEU)²⁵. In accordance with this principle, regulatory measures that restrict economic freedoms and fundamental rights may be implemented at an early stage if an evaluation concludes that a risk is likely to evolve into a danger that could cause serious damage, particularly in relation to human life and health. In the EU context, risk-based AI regulation is therefore closely connected to the *precautionary principle*.²⁶ This principle allows state authorities to impose restrictions upon technologies or activities, if a technology or behaviour is deemed to be highly risky. Even in the absence of certainty regarding the potential for damage to occur, due to a lack of appropriate knowledge about the full extent of the risks involved, the freedom to develop new technologies and to commercialise them may already be subject to pre-emptive restrictions in the interest of risk prevention. That said, the AI Act in its implementation phase seems to be increasingly under pressure not to excessively constrain market actors in the

²⁰ Martin Ebers, Truly Risk-based Regulation of Artificial Intelligence How to Implement the EU’s AI Act, in: European Journal of Risk Regulation 16 (2025), pp. 684-703.

²¹ Gavin Sullivan/Dimitri Van Den Meerssche, The Legal Infrastructures of UK Border Control—Cerberus and the *Dispositif* of Speculative Suspicion, in: German Law Journal 25 (2024), pp. 1308–1342 (1310 f.); Louise Amoore, The Deep Border, in: Political Geography 109 (2024), pp. 1-9 (1, 5 f.).

²² Ibid.

²³ Giovanni De Gregorio/Pietro Dunn, The European risk-based approaches: Connecting constitutional dots in the digital age, in: Common Market Law Review 59 (2022), pp. 473-500 (496 ff.).

²⁴ *Jale Tosun*, How the EU Handles Uncertain Risks: Understanding the Role of the Precautionary Principle, in: Journal of European Public Policy 20 (2013), pp. 1517–1528; *David Vogel*, Trading up: Consumer and Environmental Regulation in a Global Economy, 1997; *David Vogel*, Trading up and Governing across: Transnational Governance and Environmental Protection, in: Journal of European Public Policy 4 (1997), pp. 556-571 (557 ff.).

²⁵ Consolidated versions of the Treaty on European Union and the Treaty on the Functioning of the European Union of 26 October 2012, OJ C 326.

²⁶ *Smuha/Yeung* (fn. 10), p. 232. See also, *Samantha Besson*, *La due diligence* en droit international, in: Recueil des cours de l’Académie internationale de La Haye 409 (2020), pp. 153-398, (334, 322 ff.).

name of fostering innovation,²⁷ even if the baseline for assessing what ‘excess’ means in that context appears to be oftentimes missing.²⁸

2. The risk-based approach

In the context of AI regulation, the different concepts of risk-based, principle-based, precautionary principle-based, and rights-based approaches all play a role. This is due to the socio-technical dimensions of AI, where a broad risk-based approach is necessary, considering both individual and societal risks. This approach needs to be complemented by a precautionary principle-based approach, where unacceptable risks need to be defined. These approaches, however, are not uniformly defined and applicable. Different digital regulations follow different concepts. The EU’s General Data Protection Regulation (GDPR)²⁹, for instance, can be described as a bottom-up, risk-based approach, while the AI Act can be considered a top-down approach, and the Digital Ser-

vices Act (DSA)³⁰ contains both perspectives.³¹ The situation is rendered more complex by the fact that a rights-based approach is also being partially pursued in the AI Act.

Thus, considering fundamental rights violations and measuring threats and impacts on them, based on methods that are imposed by some new and existing legislation, has become a significant dimension in digital regulation. However, despite the explicit call in many regulations to consider fundamental rights impacts, opposing views are fundamentally at odds with this. On the one hand, scholars in business and economics maintain that virtually any phenomenon can be quantified; on the other hand, human rights scholars emphasize the non-discretionary, intrinsic nature of rights such as human dignity, arguing that these values resist measurement altogether.³² One goal of the new digital legislation attempts mentioned is to bring these views together. This can be achieved by analysing how to measure potential infringements of fundamental rights from an *ex ante* and *ex post* viewpoints. This means, on one side, predicting and quantifying the potential severity of human rights risks before they manifest, through *ex ante* assessments, and also defining the consequences of risk realisation in order to mitigate them (precautionary approach). On the other side, *ex post* assessments are usually realised through courts of law to clarify whether a matter that has already been concluded constitutes a human rights violation. In such cases, it is

²⁷ For further discussion on this topic, see: *De Gregorio/Dunn* (fn. 23), pp. 477 f.; European Digital Rights et al. (fn. 6).

²⁸ See for the recent developments in that regard: European Commission (fn. 5). See, for example, the criticisms addressed at the methods of the European Commission and its lack of an evidence-based approach for its proposals: *René Mahieu*, The Ominous Omnibus: Dismantling the Right of Access to Personal Data, *Verfassungsblog* of 3 December 2025, available at: <https://verfassungsblog.de/digital-omnibus-right-of-access-to-personal-data/> (last visited 3 December 2025); *Itxaso Domínguez De Olazábal*, The EU’s Digital Omnibus Must Be Rejected by Lawmakers. Here is Why, *Tech Policy Press* of 3 December 2025, available at: <https://www.techpolicy.press/the-eus-digital-omnibus-must-be-rejected-by-lawmakers-here-is-why/> (last visited 3 December 2025).

²⁹ Regulation (EU) 2016/679 of 27 April 2016, OJ L 119.

³⁰ Regulation (EU) 2022/2065 of 19 October 2022, OJ L 277.

³¹ *De Gregorio/Dunn* (fn. 23), pp. 477 f.

³² See *Gianclaudio Malgieri/Cristiana Santos*, Assessing the (severity of) impacts on fundamental rights, in: *Computer Law & Security Review* 56 (2025), pp. 1-18 (1 f.).

usually determined whether the *ex ante* measures, like risk management systems, have been properly in place. Combining *ex ante* (regulatory compliance) and *ex post* (judicial remedies) measures leads to a comprehensive approach safeguarding fundamental rights in AI regulation.³³

The developments in EU digital legislation are shifting in that regard, seeking so-called ‘optimal precaution’ in specific contexts are considered more suitable than a pure maximalist precautionary approach, in the sense of minimising risks at all costs through imposing maximum precaution.³⁴ In addition, rights-based approaches are also integrated and can foster the safeguarding of human rights if the precautionary approach is able to consider fundamental rights as a form of normative uncertainty (which, naturally, imposes limitations).³⁵ Thus, the different approaches not only coexist in EU digital legislation but are also mutually dependent.

a) Critique of the risk-based approach

A criticism directed to the risk-based approach which involves the determination of the scope or scale of a concrete situation or a perceived threat, contends that it is useful only in technical environments.³⁶ In such situations, companies evaluate their own operational risk. The AI Act’s rules concerning, for example, the notifying bodies foreseen in the Act, seem to be heading in this direction. What are these

risks weighed against? The selected approach would have companies evaluate their operational risk against people’s fundamental rights. However, from a human rights perspective, we disagree with this interpretation. Human rights, at their core, cannot be weighed against companies’ interests and must be guaranteed regardless of a risk category based on external considerations.³⁷ Regardless of potential business gains, businesses have a responsibility to respect human rights and avoid causing or contributing to human rights abuses.³⁸

To fully grasp the complexities of this interaction, it is essential to recognise that the methods used for identifying risks or assessing protected rights differ significantly. Risks to people cannot be easily integrated into corporate risk matrices, as the criteria for prioritization are distinct.³⁹

³³ Ibid.

³⁴ De Gregorio/Dunn (fn. 23), p. 478.

³⁵ Malgieri/Santos (fn. 32), p. 5.

³⁶ Fanny Hidvegi/Daniel Leufer/Estelle Massé, The EU Should Regulate AI on the Basis of Rights, Not Risks, Access Now of 17 February 2021, available at: <https://www.accessnow.org/eu-regulation-ai-risk-based-approach/> (last visited 21 August 2025).

³⁷ Ibid. See also for the differences of product safety regulation and human rights protection within the AI Act and the challenges this presents: Marco Almada/Nicolas Petit, The EU AI Act: Between the rock of product safety and the hard place of fundamental rights’, in: Common Market Law Review 62 (2025), pp. 85–120. Counterpoint: economic interests can be weighed against human rights protection, as in the ECtHR, *López Ostra v. Spain* (16798/90), judgment of 9 December 1994, para. 58, in which the Court considered “that the State did not succeed in striking a fair balance between the interest of the town’s economic well-being” and “the applicant’s effective enjoyment of her right to respect for her home and her private and family life”. Furthermore, private corporations do play an important role in implementing human rights at the international level, taking also their commercial activities into account. We maintain that the logic of fundamental rights risk cannot be incorporated to a commercial risk-based approach, because they do not pursue compatible objectives.

³⁸ UN Guiding Principles on Business and Human Rights, UN Doc. A/HRC/17/31, Annex, Chapter II.

³⁹ Malcolm Rog, Corporate Human Rights Due Diligence, Harvard Kennedy School, Working Paper No. 81 of December 2022 available at:

Furthermore, as analysed above, a purely risk-based approach rather than a rights-based approach is generally inappropriate to protect fundamental rights, as it fails to address the non-negotiable minimum core of human rights.⁴⁰ As already stated, human rights cannot be measured or quantified on a scale from trivial to severe. However, the manner in which the AI Act imposes a fundamental rights impact assessment (Art. 27 AI Act) suggests the opposite. The concept of human rights is, by contrast, based on a binary logic, whereby an act is either legal or illegal. It follows that the AI Act might fail in its own ambition to safeguard fundamental rights at this point. The mutual dependence on risk- and rights-based approaches in regulating AI should be given greater focus in the future.

b) The risk-based approach in the context of law enforcement

In this regard, the area of law enforcement is in a particular state of tension.⁴¹ Law enforcement authorities, decision-makers and society at large can be perceived as needing to achieve greater public safety. One potential method of achieving this might be to enhance the technological capacities of law enforcement agencies, for example by improving their ability to handle large volumes of data.⁴² Therefore, some argue that if the use of AI-based

systems can improve security, it may be in the interest of the state and society to use such systems.⁴³ Apart from that, it is one of the areas falling within the scope of the AI Act that poses great ethical and fundamental rights risks, alongside with the highly related field of the use of AI in migration, asylum and border control management. Thus, high standards and consistent rules must be established for the use of AI applications in law enforcement. To some extent this tension between enhancing technological capacities and addressing great challenges for ethics and fundamental rights can be analysed in the final version of the AI Act, by analysing the newly introduced category of ‘sensitive operational data’, which refers to: “operational data related to activities of prevention, detection, investigation or prosecution of criminal offences, the disclosure of which could jeopardise the integrity of criminal proceedings” (Art. 3 para. 38 AI Act).⁴⁴ The issue with this definition is that it offers too much interpretational leeway to law enforcement agencies acting as deployers, which may systematically endanger human rights protection by creating a loophole. While

https://www.hks.harvard.edu/sites/default/files/centers/mrcbg/files/CRI_81_AWP_FINAL.pdf (last visited 21 August 2025), pp. 68 ff.

⁴⁰ Ebers (fn. 20), pp. 689 ff.

⁴¹ For a comprehensive overview of the area of law enforcement in a risk-based approach, see: Steven Kleemann/Hartmut Aden, *Governing High Risk Artificial Intelligence for Law Enforcement: Strengths and Weaknesses of the European Union’s Risk-Based Approach*, in: *European Journal of Policing Studies* (forthcoming).

⁴² See for instance, European Commission, *Communication from the Commission to the European*

Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, *Roadmap for lawful and effective access to data for law enforcement*, COM(2025) 349 final of 24 June 2025, pp. 12 ff.; Dean Wilson, *Policing*, in: Mareile Kaufmann/Heidi Mork Lomell (ed.), *De Gruyter Handbook of Digital Criminology*, 2025, pp. 363–370 (368). *Contra*, Raphaël Challier/Myrtille Picaud/Florent Castagnino, *De la « safe city » aux dispositifs numériques de sécurité urbaine*, in: *Réseaux* 251 (2025), pp. 11–43 (25 f., 30).

⁴³ See: Yasmine Ezzeddine/Petra Saskia Bayer/Helen Gibson, *Safety, Privacy, or Both: Evaluating Citizens’ Perspectives around Artificial Intelligence Use by Police Forces*, in: *Policing & Society* 33 (2023), pp. 861–876 (862 f.).

⁴⁴ The concept of “sensitive operational data” can be found in various places in the regulation, such as in Art. 5 paras. 4 and 7, Art. 26 paras. 5 and 10, Art. 46 para. 3 AI Act.

other regulations, such as the GDPR for instance, define “special categories of personal data” in Art. 9 GDPR, there appears to be no further specifications required for this newly introduced category of data in law enforcement, which leaves far too much room for interpretation.

With regard to the risk classification, Art. 5 AI Act defines “prohibited artificial intelligence practices” as the highest risk category.⁴⁵ As broad exceptions will be allowed, the term is misleading.⁴⁶ This applies in particular to the prohibition of biometric facial recognition in public spaces that foresees several exceptions for law enforcement.⁴⁷ Furthermore, the use of so-called ‘*ex post*’ remote biometric identification in public spaces is authorised as a non-real-time use of biometric identification. However, this is only permitted for law enforcement under the conditions set out by the AI Act when the AI system is classified as high-risk.⁴⁸ These use cases must therefore be documented in police files and made available to the supervisory authorities upon request, as well as being reported annually to those authorities (see Art. 26 para. 10 AI Act). There is a very thin line between so-called ‘prohibited’ real-

time remote identification and its *ex post* forms, which can endanger the protection of fundamental rights if they are not strictly defined and controlled.⁴⁹

During the negotiations, the European Commission’s Proposal of the AI Act appears to have bowed to pressure from law enforcement representatives⁵⁰ and some EU Member States.⁵¹ In a resolution on AI in criminal law, the European Parliament drew attention to the relevance of this conflict.⁵² It also took up this issue in its compromise proposal in favour of a general ban, with no exceptions for law enforcement agencies, on the use of ‘real-time’ remote biometric identification systems in publicly accessible spaces.⁵³

⁴⁵ For a detailed overview of prohibited AI practices, see: European Commission (fn. 11).

⁴⁶ See: *Dimitrios Linardatos*, Auf dem Weg zu einer Europäischen KI-Verordnung – Ein (kritischer) Blick auf den aktuellen Kommissionsentwurf, in: *Zeitschrift für das Privatrecht der Europäischen Union* 19 (2022), pp. 58–70 (60); *Michael Veale/Frederik J. Zuiderveen Borgesius*, Demystifying the Draft EU Artificial Intelligence Act – Analysing the Good, the Bad, and the Unclear Elements of the Proposed Approach, in: *Computer Law Review International* 22 (2021), pp. 97–112 (101); *Andreas Ebert/Indra Spiecker gen. Döhmann*, Der Kommissionsentwurf für eine KI-Verordnung der EU, in: *Neue Zeitschrift Für Verwaltungsrecht* 6 (2021), pp. 1188–1893 (1189 f.).

⁴⁷ European Commission (fn. 11), para. 294.

⁴⁸ *Ibid.*, paras. 427 f.

⁴⁹ *Daragh Murray*, Police Use of Retrospective Facial Recognition Technology A Step Change in Surveillance Capability Necessitating an Evolution of the Human Rights Law Framework, in: *Modern Law Review* 87 (2024), pp. 833–863 (837); European Commission (fn. 11), para. 310; *Eric Töpfer/Steven Kleemann*, Polizeiliche Gesichtserkennung – Menschenrechtliche Herausforderungen einer Risikotechnologie, *Deutsches Institut für Menschenrechte* of August 2025, available at: https://www.institut-fuer-menschenrechte.de/fileadmin/Redaktion/Publikationen/Analyse_Studie/Analyse_Polizeiliche_Gesichtserkennung_01.pdf (last visited 27 October 2025).

⁵⁰ *Veale/Borgesius* (fn. 46), p. 98; Access Now, Europe’s Approach to Artificial Intelligence: How AI Strategy is Evolving, December 2020, available at: <https://perma.cc/X3JM-2M6A> (last visited 21 August 2025).

⁵¹ See *Maggiore/Miñano/Schumann* (fn. 6).

⁵² European Parliament, Artificial Intelligence in Criminal Law and Its Use by the Police and Judicial Authorities in Criminal Matters, P9_TA(2021)0405 of 6 October 2021, available at: https://www.europarl.europa.eu/doceo/document/TA-9-2021-0405_EN.pdf (last visited 21 August 2025).

⁵³ European Parliament, DRAFT Compromise Amendments on the Draft Report Proposal for a Regulation of the European Parliament and of the Council on Harmonised Rules on Artificial Intelli-

Several facial recognition applications for law enforcement purposes and the various scenarios in which they could be used are conceivable⁵⁴ – in addition to the possibility for law enforcement authorities to use so-called ‘remote biometric identification systems’ if they satisfy the conditions laid out in the Act.⁵⁵ Despite protests from civil society organisations against the exceptions, which echoed similar criticisms from the EU’s data protection authorities and the European Parliament, the final agreement does not include a real ban.⁵⁶ A step forward safeguarding fundamental rights – though it must be closely monitored –, is the Council of Europe’s so-called AI Framework Convention.⁵⁷ Unlike the AI Act, which also seeks to harmonise economic interests, this Framework Convention is primarily concerned with the protection of human rights. However, the extent to which it fulfils its stated objective of protecting fundamental rights remains to be determined, as it has a more limited ambition than the AI Act.⁵⁸

gence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM(2021)0206 – C9 0146/2021 – 2021/0106(COD) of 9 May 2023.

⁵⁴ *Töpfer/Kleemann* (fn. 49).

⁵⁵ *Veale/Borgesius* (fn. 46), pp. 101 f.

⁵⁶ For a comprehensive overview of the debate on bans, see *Catharina Rudschies/Ingrid Schneider*, *The Long and Winding Road to Bans for Artificial Intelligence: From Public Pressure and Regulatory Initiatives to the EU AI Act*, in: *Digital Society 4* (2025), pp. 1-27.

⁵⁷ Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law of 05 September 2024, CETS No. 225.

⁵⁸ *Francesco Paolo Levantino/Frederica Paolucci*, *Advancing the Protection of Fundamental Rights Through AI Regulation: How the EU and the Council of Europe are Shaping the Future*, in: Philip Czech et al. (ed.), *European Yearbook on Human Rights*, 2024, pp. 3-37 (11 f.); European Data Protection Supervisor, *Opinion 20/2022 on the Recommendation for a Council Decision authorising*

The AI Act generally classifies the use of AI systems by law enforcement agencies as high-risk (Recitals 59-60 AI Act). It remains an open question how the risk categories should be applied to AI systems by law enforcement agencies for purposes not listed in the aforementioned Annex III. To what extent does the AI Act regulate law enforcement per se? Some European legislators have successfully argued that the application of the proposed requirements of the AI Act should be excluded precisely in those contexts where the threats to fundamental rights are the greatest: national security, defence, transnational law enforcement, as well as research and development.⁵⁹ This exclusion of the material scope of the AI Act was justified by the argument that national security is generally excluded from the scope of EU law and that, according to Recital 24 AI Act, public international law would be “the more appropriate legal framework for the regulation of AI systems in the context of the use of lethal force and other AI systems in the context of military and defence activities”. In principle, this exclusion depends exclusively “on the purposes of the AI system, not the entities carrying out the activities with that system.” However, such AI systems “must be placed on the market, put into service or used exclusively for military, defence or national security pur-

the opening of negotiations on behalf of the European Union for a Council of Europe convention on artificial intelligence, human rights, democracy and the rule of law, 13 October 2022.

⁵⁹ *Douwe Korff*, *Opinion on the Implications of the Exclusion from New Binding European Instruments on the Use of AI in Military, National Security and Transnational Law Enforcement Contexts*, 2022, available at: https://ecnl.org/sites/default/files/2022-10/ECNL%20Opinion%20AI%20national%20security_0.pdf (last visited 27 October 2025), p. 28; *Smuha/Yeung* (fn. 10), p. 235.

poses”.⁶⁰ However, the applicable criteria can endanger fundamental rights, as the distinction between AI systems that are and are not excluded from the scope of the AI Act does not appear to be consistent in practical terms.⁶¹ The European Commission’s Guidelines state that dual-use AI systems—those intended for both civilian and security purposes—are covered by the AI Act.⁶² However, this does not limit national security, defence, or military actors from using such systems for those specific purposes, regardless of the entity’s nature.⁶³ Agencies like Europol, Frontex, and national police forces may operate outside the AI Act (and the GDPR) when acting under other legal instruments,⁶⁴ and large EU IT systems (Eurodac, SIS, ETIAS) are only subject to the AI Act after 2 August 2030 (Art. 111, Annex X AI Act). Article 2 para. 4 AI Act also excludes third country public authorities or international organisations using AI in EU-linked law enforcement or judicial cooperation, a provision that can be extended to private contractors when they are acting on behalf of those authorities.⁶⁵ This raises questions about cases such as EncroChat, where law enforcement, intelligence services, and private firms collaborated to breach encrypted communications.⁶⁶ The exemption applies only if the cooperation framework

includes adequate safeguards for fundamental rights, overseen by the relevant market surveillance authorities.⁶⁷

Therefore, despite the added clarifications regarding the scope of application of the AI systems that are excluded from the scope of the AI Act, there are concerns that some AI systems might still be used by security or judicial actors without respecting the obligations normally applicable to law enforcement. Only time will tell if the AI Act’s scope of application might be more or less protective in comparison with the logic that has led to the adoption of the GDPR and the so-called Law Enforcement Directive (LED).⁶⁸

c) Filtering fundamental rights protection and broadening the scope for avoiding high-risk classification

The Commission’s Proposal for the AI Act was logical in that all high-risk applications in Annex III would have to comply with certain obligations. However, due to industry and state interventions an additional ‘filter’ was integrated. This ‘filter’⁶⁹ can be found in Art. 6 para. 3 AI Act and states that AI systems that are intended

Crime, Criminal Law and Criminal Justice 30 (2022), pp. 309-328.

⁶⁰ European Commission (fn. 11), para. 24.

⁶¹ See, *Plixavra Vogiatzoglou*, The AI Act National Security Exception: room for manoeuvres?, *Verfassungsblog* of 9 December 2024, available at: <https://verfassungsblog.de/the-ai-act-national-security-exception/> (last visited 15 August 2025).

⁶² *Ibid.*

⁶³ *Ibid.*, para. 25.

⁶⁴ *Korff* (fn. 59), p. 29.

⁶⁵ European Commission (fn. 11), para. 29.

⁶⁶ *Jan-Jaap Oerlemans/Dave van Toor*, Legal Aspects of the EncroChat Operation: A Human Rights Perspective, in: *European Journal of*

⁶⁷ European Commission (fn. 11), para. 29; See also Recital 22 AI Act.

⁶⁸ Directive (EU) 2016/680 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA of 27 April 2016.

⁶⁹ The so-called filter provision was introduced under the influence of the Council of the EU and the European Parliament during the AI Act negotiations of the AI Act, which concluded with the trilogue, see also: *Palmiotto* (fn. 6) pp. 780 f., 787 f.; *Stewart* (fn. 14).

for a narrow procedural task, such as confirming or improving an accessory factor of a human assessment or performing a preparatory task, can be exempted from categorisation as high-risk systems under certain conditions. Moreover, this new feature was integrated despite a counter-mobilisation of civil society organisations, a critical letter on this very issue from the UN High Commissioner for Human Rights⁷⁰ and a damning negative opinion from the European Parliament’s legal service.⁷¹

The introduction of a structural loophole now seems to broaden the remit of the existing high-risk classification, already open to criticism. Despite the introduction of an *ex post* corrective mechanism in the powers attributed to national market surveillance authorities for controlling and sanctioning that a provider has wrongly classified an AI system as non-high-risk according to Art. 80 AI Act, the self-regulatory powers of AI providers organised under the AI Act can be particularly dangerous in the fields of law enforcement and migration.⁷²

d) Drawbacks of the risk-based approach

Lastly, the risk-based approach also contains weaknesses. The current categorisa-

tion and the systems listed in Annex III (high-risk systems) need to be differentiated in some respects. The listed risks are classified according to external considerations rather than the legal interests involved, such as law enforcement or critical infrastructure on one side and potentially endangered fundamental rights on the other. The classification itself is based on the idea, that these areas are of particular relevance, which is true for a rough template, but it is not conclusive or sufficient to ensure comprehensive protection of fundamental rights.⁷³ Thus, the current risk classification, especially the high-risk category, needs to be conceptualised and implemented in a more nuanced way. In this context, risk classification should not be limited to the three broad categories of high-, medium- and low-risk AI systems, in addition to the prohibited AI systems: within these categories, gradations between different sub-levels of risk would facilitate a more differentiated risk assessment. In the field of law enforcement, which is generally and rightly considered in the high-risk category, a gradual distinction from ‘low-high-risk’ to ‘high-high-risk’ should be introduced. The dangers posed by such systems for fundamental rights vary. Consequently, there is a need for an interplay of regulatory approaches without creating gaps or unclear risks in the application of AI.⁷⁴

⁷⁰ United Nations, Open Letter from the United Nations High Commissioner for Human Rights to European Union institutions on the European Union Artificial Intelligence Act (“AI Act”) of 8 November 2023, available at: https://www.ohchr.org/sites/default/files/2024-12/Tu%CC%88rk_open_letter_European_Union_highlights_issues_with_AI_Act_8_11_23.pdf (last visited 23 August 2025).

⁷¹ Daniel Leufer/Caterina Rodelli/Fanny Hidvegi, Human Rights Protections... with Exceptions, Access Now of 14 December 2023, available at: <https://www.accessnow.org/whats-not-in-the-eu-ai-act-deal/> (last visited 21 August 2025).

⁷² Stewart (fn. 14), pp. 129 f.

⁷³ Hannah Ruschemeier, AI as a challenge for legal regulation – the scope of application of the artificial intelligence act proposal, in: ERA Forum 23 (2023), pp. 361–376 (373).

⁷⁴ For an attempt of a more nuanced risk classification of the area of law enforcement, see: Klee-mann/Aden (fn. 41).

3. Direct responsibility and accountability mechanisms: late additions to the AI Act

The original proposal of the AI Act did not include a mechanism for individuals who are harmed or otherwise negatively affected by AI systems to file a complaint or seek redress.⁷⁵ In the final version, however, a new Section 4 (Remedies) in Chapter VII has been added, comprising Articles 85 and 86 AI Act. It is already foreseeable that these complaints mechanisms will serve as a channel between AI developers, deployers, users and those affected by AI-based decisions. This feature of the institutional architecture for the implementation of the AI Act is of fundamental importance in terms of AI accountability and responsibility, as Art. 85 AI Act provides the main remedy directly available to lay persons affected by AI systems under the scope of the Act, the right to lodge a complaint with the competent national supervisory authority, “[w]ithout prejudice to other administrative or judicial remedies” (Art. 85 para. 2 AI Act). The right to lodge a complaint is widely accessible as it addresses “any natural or legal persons having grounds to consider that there has been an infringement”, opening the door to the possibility of initiating collective

forms of legal action on the basis of the AI Act.⁷⁶

Furthermore, according to Art. 86 para. 1 AI Act, “[a]ny affected person subject to a decision which is taken by the deployer on the basis of the output from a high-risk AI system [...] and which produces legal effects or similarly significantly affects that person in a way that they consider to have an adverse impact on their health, safety or fundamental rights shall have the right to obtain from the deployer clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken”. Art. 86 para. 2 AI Act contains some exceptions to this right to an explanation. The extent to which this provision can provide actual protection against a decision taken by a law enforcement authority requires further analysis. Furthermore, the role of national courts and domestic (constitutional) law will be of utmost importance, as fundamental rights violations by national law enforcement agencies will most likely be heard there first.

The analysis of the AI Act’s potential for protecting fundamental rights in the context of using AI tools for law enforcement purposes requires examining the institutional architecture that the European Union and its member states are progressively establishing. Accountability and responsibility mechanisms will develop within this complex multi-level institutional environment, which in practice will affect the scope of protection theoretically

⁷⁵ Palmiotto (fn. 6), pp. 778 f.; European Digital Rights, The EU AI Act and fundamental rights: Updates on the political process, EDRI of 9 March 2022, available at: <https://edri.org/our-work/the-eu-ai-act-and-fundamental-rights-updates-on-the-political-process/> (last visited 21 August 2025); European Digital Rights, Civil society calls on the EU to put fundamental rights first in the AI Act, EDRI of 30 November 2021, available at: <https://edri.org/our-work/civil-society-calls-on-the-eu-to-put-fundamental-rights-first-in-the-ai-act/> (last visited 21 August 2025); Veale/Borgesius (fn. 46), p. 111.

⁷⁶ European Center for Not-for-Profit Law, Towards an AI Act that serves people and society: Strategic actions for civil society and funders on the enforcement of the EU AI Act, ECNL of August 2024, available at: https://ecnll.org/sites/default/files/2024-09/AIFUND_ECNL_AI_ACT_Enforcement_2024.pdf, (last visited 21 August 2025), p. 39.

offered to affected persons. Given the nature of the challenges in these sensitive fields of AI application—where the AI Act provides numerous exceptions or ‘backdoors’ benefiting public security agencies using AI—we assess how AI contestability could legally, formally, or informally, and spontaneously arise in response to these issues.

III. Accountability Mechanisms under the AI Act and AI contestability

Responsibility and accountability under the AI Act require not only an institutional architecture and oversight authorities that can effectively monitor and sanction compliance with its obligations, but also effective substantive rights that at least enable affected persons and laypersons to contest AI-based decisions, particularly those involving the use of AI for law enforcement purposes. In this regard, NGOs have criticised the right to lodge a complaint and Section 5 in Chapter IX on remedies in the AI Act for lacking teeth, stating that “it remains unclear how effectively these [supervisory] authorities will be able to enforce compliance and hold violators accountable”.⁷⁷ Regarding the future implementation of remedies, the multiplicity

of oversight bodies may further weaken the effectiveness of legal means of contestation of AI-based decisions, activities and processes. According to expert consultations organised within the framework of ‘Accountability Principles for Artificial Intelligence (AP4AI) in the internal security domain’, the “[principle of enforceability] requires that relevant oversight bodies and enforcement authorities have the necessary power and means to respond appropriately to instances of non-compliance with applicable obligations by those deploying AI in a criminal justice context”.⁷⁸

This so-called ‘many eyes phenomenon’⁷⁹ in the context of AI regulation⁸⁰ needs to be considered, with respect to the complex institutional architecture set up by the AI Act.⁸¹ This simply means that the effective implementation of responsibility and accountability in relation to AI and its use for law enforcement purposes may be seriously hampered by the multiplicity of European and national authorities that will be responsible for controlling compliance with the requirements of the AI Act. This illustrates the challenges that lie

⁷⁷ European Digital Rights and AI coalition partners, EU’s AI Act fails to set gold standard for human rights, EDRI of 3 April 2024, pp. 3 f., available at: <https://edri.org/our-work/eu-ai-act-fails-to-set-gold-standard-for-human-rights/> (last visited 21 August 2025); Access Now, The EU AI Act: a failure for human rights, a victory for industry and law enforcement, 13 March 2024, available at: <https://www.accessnow.org/press-release/ai-act-failure-for-human-rights-victory-for-industry-and-law-enforcement/> (last visited 21 August 2025).

⁷⁸ *Babak Akhgar et al.*, Accountability Principles for Artificial Intelligence (AP4AI) in the Internal Security Domain, AP4AI Framework Blueprint. Accountability Principles for Artificial Intelligence (AP4AI), 2022, p. 38.

⁷⁹ *Mark Bovens*, Analyzing and Assessing Accountability: A Conceptual Framework, in: *European Law Journal* 13 (2007), pp. 447–468 (455 ff.).

⁸⁰ *Claudio Novelli/Mariarosaria Taddeo/Luciano Floridi*, Accountability in artificial intelligence: what it is and how it works, in: *AI & Society* 39 (2024), pp. 1871–1882 (1875).

⁸¹ See also for this issue: *Sol Martinez Demarco/Milan Tahraoui/Steven Kleemann*, The siloed logic and the implementation of the Artificial Intelligence Act in the law enforcement context: legal and ethical analysis of the applicability of accountability and responsibility to high-risk AI systems, *Routledge Studies in Surveillance* (forthcoming).

ahead in creating and embedding concrete mechanisms of accountability and responsibility in the practice of sensitive areas of law enforcement activities.

Firstly, we refer to the two main types of supervisory bodies qualified by the AI Act as the Member States' national competent authorities. We briefly introduce the so-called market surveillance authorities and the notifying authorities, as well as notified bodies for conducting conformity assessments. As we later explain, although the term 'national competent authorities' is generic, it actually adds another layer to the 'many eyes' problem. Indeed, this term is used in different ways, referring either to (i) data protection authorities with additional tasks, competences and means, (ii) newly established bodies responsible for implementing the Act at the national level, or (iii) to several other possible competent independent national public authorities, if needed.⁸² Finally, we will frame the interests of considering AI contestability by examining the possible existence of a 'right to contest' in the context of AI, also distinguishing between corrective and non-corrective forms of contestability. The argument is that AI contestability is of critical importance, as the AI Act has significant shortcomings regarding accountability and responsibility for the use of AI for public security purposes.

1. Insights on the AI Act Complex Governance Architecture and its Role for AI Accountability and Responsibility

With regard to accountability and responsibility for the use of AI in law enforcement, as well as migration, asylum and

border management control purposes, two key institutions established by the AI Act are the market surveillance authorities and the notifying bodies, which will act as national competent authorities (Art. 70 para. 1 AI Act). In particular, they will be responsible for ensuring that the use of AI systems for public security purposes does not compromise the health, safety and fundamental rights,⁸³ even if the AI Act also confers powers to other institutions for enforcing the AI Act in line with fundamental rights.

Two important powers given to national authorities to ensure that high-risk AI systems in law enforcement comply with the AI Act are notable. First, Article 74 para. 2 AI Act states that these authorities should have full access to the documentation and datasets used for developing such systems, including training, validation, and testing data. This access can be provided through APIs or other secure remote access methods, as long as it is necessary for their tasks. Second, the AI Act stipulates that market surveillance authorities overseeing high-risk AI systems in biometrics—when used for law enforcement, migration, asylum, border control, justice administration, or democratic processes—should have strong investigative and corrective powers (see also Recital 159 AI Act). These include the ability to access all personal data being processed and any other information needed to carry out their duties. One exception to these powers involves the previously mentioned category of 'sensitive operational data'.⁸⁴

The development of a complex institutional architecture for the implementation

⁸² See below, for example, the independent public authorities that France has designated on the basis of Art. 77 AI Act.

⁸³ *Marion Ho-Dac*, La protection des droits fondamentaux dans l'AI Act : Essai de cartographie critique, in : RTD. Euro. 2025, pp. 615-633.

⁸⁴ See Section II. 2.

of the AI Act has been informed by debates and criticism raised about serious enforcement issues that affect the implementation of other, already-established regulations, such as the GDPR, also leading to a deficit in terms of responsibility and accountability.⁸⁵ One frequent critique concerns the difficulties to enforce the GDPR across the EU, as the EU Member States' national data protection authorities have divergent legal and political preferences.⁸⁶ Indeed, significant challenges lie ahead regarding accountability and responsibility for the development and deployment of AI systems in law enforcement contexts. This is particularly pertinent given that EU Member States have demonstrated a propensity to advocate for a reduced set of obligations.⁸⁷ Consequently, such preferences may influence the institutional framework of oversight

mechanisms within law enforcement domains, which are predominantly shaped by the decisions of EU Member States. 'Excessive accountability'⁸⁸ is an interesting concept for criticising this complex architecture from the viewpoint of fundamental rights protection, as it describes the accumulation and network of accountability mechanisms that have produced negative side effects in terms of increasing costs, red tape, and a deterioration of public values such as effectiveness, efficiency, trust, and learning.⁸⁹

2. AI Contestability

As the rapid diffusion of AI in the fields of law enforcement activity and beyond is strongly pushed by states' authorities and dominant firms, without much active role conferred to affected persons and consideration for the broad public,⁹⁰ the

⁸⁵ In this sense, see for example *Yiran Lin*, More Than an Enforcement Problem: The General Data Protection Regulation, Legal Fragmentation, and Transnational Data Governance, in: *Columbia Journal of Transnational Law* 62 (2024), pp. 1-39 (20 ff.).

⁸⁶ See for instance, *Giulia Gentile/Orla Lynskey*, Deficient by Design? The Transnational Enforcement of the GDPR, in: *International and Comparative Law Quarterly* 71 (2022), pp. 799-830 (818). *Contra*, a member of the European Parliament involved in the legislative process for this Act, claims that the AI Act will achieve a better level of enforcement due to the quality of the future market surveillance authorities. See, *Laura Caroli*, Will the EU AI Act work? Lessons learned from past legislative initiatives, future challenges, IAPP News of 17 April 2024, available at: <https://iapp.org/news/a/will-the-eu-ai-act-work-lessons-learned-from-past-legislative-initiatives-future-challenges> (last visited 21 August 2025).

⁸⁷ *Palmiotto* (fn. 6); Council of the European Union, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts – General approach, 2021/0106(COD), 14336/22 of 11 November 2022, para. 4.2.

⁸⁸ *Mark Bovens/Thomas Schillemans*, Meaningful Accountability, in: *Mark Bovens/Robert Goodin/Thomas Schillemans* (ed.), *The Oxford Handbook of Public Accountability*, 2014, pp. 673-682 (674).

⁸⁹ Remarkably, the proposed Digital Omnibus on AI (European Commission (fn. 5)) aims to "simplify" the supervisory architecture primarily by granting more centralized powers to the AI Office for regulating general-purpose AI models with broad EU-wide impact (pp. 27 f.). Amendments are also envisaged for the supervision of fundamental rights by national authorities and their cooperation, but to a much lesser extent (pp. 29 f.). Essentially, the proposal will further reduce the scope of AI-based activities that can be supervised by reducing applicable obligations for high-risk systems or delaying the applicability of parts of the AI Act (pp. 2, 21 ff.).

⁹⁰ See for example, *Marie Petersmann/Dimitri Van Den Meerssche*, On phantom publics, clusters, and collectives: be(com)ing subject in algorithmic times, in: *AI & Society* 39 (2024), pp. 107-124; *Chris Jones/Romain Lanneau*, Automating Authority: Artificial Intelligence in European Police and Border Regimes, *Statewatch* of April 2025, available at:

contestability of AI development and deployment is a useful analytical concept for addressing AI accountability and responsibility. Although AI contestability is also partly embedded in international human rights law, including the right to an effective redress, it extends beyond this, encompassing social engagement practices involving AI socio-technical use-cases, even in the absence of anticipated or actual harm. If the risk-based approach had a strong bearing on the negotiation and the adoption of the AI Act, a rights-based approach can command the legal and ethical organisation and reaction to corrective and non-corrective forms of AI contestability. As the 2021 UNESCO Recommendations on the ethics of AI clearly state, it “should be recognized that AI technologies do not necessarily, per se, ensure human and environmental and ecosystem flourishing”.⁹¹ This statement reflects a human rights approach that seeks to precondition the use of AI technologies to specific justifications for its use in light of several criteria in which appropriate, proportional and legitimate aims, as well as human rights and rigorous scientific foundations play a key role.⁹² As AI can impact societies in which it is deployed far beyond the mere individual situations of persons directly and effectively affected or harmed

by a particular use case,⁹³ it is necessary to reflect on the potential contribution that a right to effective contestation of the use of AI for law enforcement purposes might bring, while contemplating the limits of a perspective arguably excessively focused on individuals⁹⁴ and on corrective forms of contestation.

According to the first paradigm, the increasing use of AI-based systems necessitates a stronger emphasis on individuals’ right to challenge decisions that affect their lives. This right arises not only from legal frameworks such as the GDPR and parts of the AI Act (Art. 85 ff. AI Act), after their introduction by amendments proposed by the European Parliament to take better into account affected persons,⁹⁵ but also from broader human rights principles and procedural guarantees enshrined in democratic legal systems. An important example is the right to an effective remedy under international human rights law⁹⁶

<https://www.statewatch.org/automating-authority-artificial-intelligence-in-european-police-and-border-regimes/> (last visited 22 August 2025); Anaëlle Beignon/Thomas Thibault/Nolwenn Maudet, Imposing AI: deceptive design patterns against sustainability, Limits ’25, 11th Workshop on Computing Within Limits, 26–27 June 2025, available at: <https://computingwithlimits.org/2025/papers/limits2025-beigon-imposing-ai.pdf> (last visited 21 August 2025).

⁹¹ UNESCO, Recommendation on the ethics of artificial intelligence, SHS/BIO/REC-AIETHICS/2021 of November 2021, para. 25.

⁹² *Ibid.*, para. 26. This paragraph stresses that: “In particular, AI systems should not be used for social scoring or mass surveillance purposes”.

⁹³ *Smuha/Yeung* (fn. 10), p. 258.

⁹⁴ *Rinaldi/Teo* (fn. 15), p. 78; European Center for Not-for-Profit Law (fn. 76), p. 39.

⁹⁵ European Parliament, Artificial Intelligence Act, Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)), P9_TA(2023)0236 of 14 June 2023, Amendments 628-630; *Palmiotto* (fn. 6), pp. 783 f.

⁹⁶ Art. 8 Universal Declaration of Human Rights of 10 December 1948, UN Doc. A/RES/217 A (III); Art. 2 para. 3 International Covenant on Civil and Political Rights of 16 December 1966, UNTS vol. 999, p. 171; Art. 13 European Convention on Human Rights of 4 November 1950, ETS No. 005; Art. 2 lit. c) Convention on the Elimination of All Forms of Discrimination against Women of 18 December 1979, UNTS vol. 1249, p. 13; Art. 6 International Convention on the Elimination of Racial Discrimination of 7 March 1966, UNTS vol. 660, p. 195.

and emerging international regulatory AI frameworks. In this sense, the contestability of AI-based decisions contributes to preserve fairness, serve justice and autonomy, while at the same time correcting errors, preventing unfair outcomes and improving transparency.

a) A right to contest and the concept of corrective contestability

An effective ‘right to contest’ certain AI-based decisions can serve as a fundamental mechanism for correcting the asymmetrical power relations created by algorithmic decision-making systems. In this sense, contestability is the ability to appeal or effectively complain about decisions made by systems, which is essential for ensuring agency and fairness in digital environments. This means that contesting a decision is not only a legal or procedural necessity, but a design concept anchored in core principles of AI regulation such as transparency, accessibility and autonomy.⁹⁷ When contestability is embedded in the design of decision-making systems, it can serve as a bridge between users, affected parties and systems, offering individuals opportunities to actively engage with and influence decisions that affect their lives.⁹⁸ This shows that this area also goes beyond individual considerations of specific legal, ethical or technical aspects and must be viewed holistically within and across individual disciplines. If a right to contest is understood and implemented in this way, it can contribute to serving principles such as fairness and justice to a greater extent and uphold constitutional values by correcting errors, and

preventing or changing unfair outcomes retrospectively. This can arguably also lead to more predictable and consistent decisions.⁹⁹

From a European legal perspective, one can refer to the GDPR (which operationalises data protection aspects of the right to privacy enshrined in Article 8 of the Charter of Fundamental Rights of the European Union)¹⁰⁰ to the case law of the European Court of Justice (ECJ) interpreting those provisions, and, where appropriate, to the OECD Council’s non-binding recommendations on AI.¹⁰¹ In its recommendations on AI, the OECD states that, in the context of transparency and explainability, it must also be possible “to provide information that enable those adversely affected by an AI system to challenge its output”.¹⁰² Although the contextual definition of the term ‘challenge’ is not explained in greater detail, the OECD’s recommendations have already shaped data protection laws around the world on many occasions in the past, and its recommendations on AI can be influential.¹⁰³ However, it is notable that the focus is on *output* and *outcome*,¹⁰⁴ rather than other

⁹⁷ Robert Patrick Collins/Johan Redström/Marco Rozendaal, The Right to Contestation: Towards Repairing Our Interactions with Algorithmic Decision Systems, in: International Journal of Design 18 (2024), pp. 95-106 (96 ff.).

⁹⁸ Ibid., pp. 97 ff.

⁹⁹ See: Margot E. Kaminski/Jennifer M. Urban, The Right to Contest AI, in: Columbia Law Review 121 (2021), pp. 1957-2048 (1974 f.).

¹⁰⁰ Charter of Fundamental Rights of the European Union of 14 December 2007, OJ C 303/1.

¹⁰¹ OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449 of 22 May 2019.

¹⁰² Ibid., p. 9.

¹⁰³ Kaminski/Urban (fn. 99), p. 1963.

¹⁰⁴ See for the latter, OECD, Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI, OECD Digital Economy Papers No. 349, February 2023, p. 32: “Users of explainable AI systems benefit from being able to understand and challenge or contest an outcome, seek redress, and earn through human-computer interfaces”.

possible subjects of contestation.¹⁰⁵ Data protection law stipulates that, pursuant to Art. 22 para. 1 GDPR, individuals must not be subject to decisions based solely on automated processing, and pursuant to Art. 22 para. 3 GDPR, the controller “shall implement suitable measures to safeguard the data subject’s rights and freedoms and legitimate interests[,] [...] at least the right [...] to *contest* the decision”. In parallel, Art. 11 LED stipulates for the area of criminal law enforcement that “a decision based solely on automated processing, including profiling, [...] [is] prohibited [...]” and that the data subject shall have at least the right to obtain human intervention. Recital 38 LED then states that “in any case, such processing should be subject to suitable safeguards, including [...] the right to obtain human intervention, [...] to obtain an explanation of the decision reached [...] or to *challenge the decision*”. Thus, it can be argued that a right to contest can exist at least against certain forms of processing of personal data, namely fully automated data processing. This is likely to apply to a significant proportion of AI-based decisions that affect individuals at an individual level. However, there is some uncertainty regarding the wording “based solely on automated processing” and the requirement that the decision “produces legal effects [...] or similarly *significantly affects*” (Art. 22 para. 1 GDPR). It is not entirely clear what thresholds apply here. Furthermore, if a right to contest is really intended to exist and be enforceable, it is odd that it is ‘hidden’ in such a place

and does not appear elsewhere in the text of the Regulation, nor is it explicitly defined or explained in more detail. Finally, the AI Act has been criticised for focusing excessively on AI providers regarding human oversight (Art. 14 AI Act), rather than on the key role of the deployers in ensuring human intervention, while simply requiring awareness of automation biases but no real obligation to act upon it.¹⁰⁶

For such a right to be effective, it would have to go beyond a mere right to rectification and at least include an obligation to examine the merits of a complaint and to give reasons for a decision, and that this right requires from the data controller to either make the automated decisions effectively contestable or to discontinue the use of the algorithmic decision-making system altogether.¹⁰⁷ Furthermore, individual rights, procedural rights and transparency rights contained in the GDPR must be taken into account in their entirety to enable an effective right to contest.¹⁰⁸ This means that, at least from the Art. 22 GDPR¹⁰⁹ in conjunction with the Arts. 13, 14, 15 GDPR¹¹⁰ and Recital 71

¹⁰⁵See however the slightly broader take on AI contestability in OECD, AI, Data Governance and Privacy: Synergies and Areas of International Co-Operation, OECD Artificial Intelligence Papers No. 22, June 2024, p. 39: “when it comes to helping persons affected by AI systems understand and contest their processes and outputs, or to help users detect algorithmic discrimination, data protection law and AI policy align”.

¹⁰⁶*Laux/Ruscheimer* (fn. 9).

¹⁰⁷*Emre Bayamlıoğlu*, The right to contest automated decisions under the General Data Protection Regulation: Beyond the so-called “right to explanation”, in: *Regulation & Governance* 16 (2022), pp. 1058-1078 (1063).

¹⁰⁸*Kaminski/Urban* (fn. 99), p. 1979.

¹⁰⁹European Data Protection Board, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, 18 December 2024, available at: https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_en (last visited 21 August 2025).

¹¹⁰*Ibid.*, para. 63.

GDPR a right to an explanation,¹¹¹ or even a right to contest, can be derived.

Moreover, the concept of a right to contest goes beyond the scope of European data protection and privacy regulations and legislative efforts. Requirements such as accountability and transparency are also found in the AI Act, which (in addition to ensuring product safety and compliance with EU law) also aims to protect fundamental rights. As outlined above, the right to contest AI-based decisions is arguably a cornerstone for ensuring fairness, justice and accountability in an increasingly automated world. Based on the GDPR, the AI Act and the broader human rights framework, a right to contest should enable individuals to challenge decisions that affect them and hold AI providers and AI deployers accountable. To be effective, challenge mechanisms must include transparency, human control and clear legal remedies so that individuals can effectively exercise their autonomy. As AI systems influence important decisions, mechanisms for contestation are indispensable tools for upholding the rule of law and addressing the ethical and societal challenges arising from algorithmic decisions. However, the difficulty lies in implementing these rights in practice, as AI systems often lack the necessary transparency to enable individuals to understand, or to challenge, their outcomes. In this regard, the ECJ has issued a remarkable landmark ruling on the transparency of algorithms, confirming the existence of a ‘right to an explanation’ in relation to automated de-

isions.¹¹² The ruling also clarified that courts and competent authorities have access to information protected by trade secrets, where necessary, to reconcile this protection with the fundamental rights of the individuals concerned – a finding that will have significant implications, particularly for organisations using high-risk AI systems in decision-making processes affecting individuals.

Thus, a right to contest, object or appeal can be derived from the aforementioned standards of the GDPR, the AI Act, the OECD recommendations and the broader human rights legal framework. In order to enable effective contestation, principles such as transparency, human intervention and explainability must be guaranteed. From a human rights perspective, this is crucial to enable individuals to effectively assert their rights, which is predicated on their awareness of algorithmic decisions or, where applicable, profiling, and their effective ability to challenge the reasons for algorithmic decisions.¹¹³ This means that transparency is particularly important. Transparency should provide context-specific and comprehensible reasons for decisions and allow for case-specific discretion.¹¹⁴ In order to achieve such transparency, it is also necessary to involve civil society actors, including those

¹¹¹Bryan Casey/Ashkon Farhangi/Roland Vogl, Rethinking Explainable Machines: The GDPR’s ‘Right to Explanation’ Debate and the Rise of Algorithmic Audits in Enterprise, in: *Berkeley Technology Law Journal* 34 (2019), pp. 143-188 (155 ff.).

¹¹²ECJ, judgment of 27 February 2025, Case C-203/22, paras. 57 ff. (Dun & Bradstreet Austria).

¹¹³See with respect to the awareness of data subjects and their reasonable expectations concerning the processing of their data, European Data Protection Board (fn. 109), paras. 93 ff.

¹¹⁴See for instance, *Rita Matulionyte*, Increasing transparency around facial recognition technologies in law enforcement: towards a model framework, in: *Information & Communications Technology Law* 33 (2024), pp. 66–84; *Evgeni Aizenberg/Jeroen van den Hoven*, Designing for human rights in AI, in: *Big Data & Society* 7 (2020), pp. 1–14 (7).

who could take on the task of challenging decisions in certain contexts, in order to develop an understanding of what constitutes a comprehensible justification and an effective challenge mechanism.¹¹⁵ Finally, transparency can also support corrective and non-corrective forms of AI contestability. Regrettably, transparency requirements have been significantly reduced during the negotiations of the AI Act in the case of high-risk AI systems used for public security purposes.¹¹⁶

It can be argued that AI contestability is not limited to the right to contest, which can be qualified as corrective AI contestability. In much of the literature dealing with AI contestability in relation to social systems theories, AI contestability is considered a means of potentially improving the development, functioning, deployment, and efficiency of AI systems or models. AI contestability is, therefore, mainly considered from the perspective of the internal logic of AI development and AI deployment to improve its use and mitigate its side effects.¹¹⁷ One example is the contestability of the acceptability of an AI system's error rate.¹¹⁸ Furthermore, these strands of literature often focus on the key role of AI developers, conceptualised as an interplay between normative principles and the translation of rules into technical design. Actors outside technical

development also tend to be considered as mere recipients or users of AI contestation. A general sensibility for ethical challenges within AI software design can be identified in the perspectives following a corrective AI contestability conception, which is supposed to lead to technology that better protects the rights, interests, and needs of affected communities and persons. Lastly, corrective AI contestability generally intervenes *ex post facto*, which complicates *ex ante* forms of contestation, such as questioning or opposing the development of an AI tool from the outset.¹¹⁹

The concept of contestation intervenes in various ways in international and regional AI regulatory frameworks, but most occurrences covered tend to be corrective in nature, as legal mechanisms are anchored in situations where individuals can contest the mere outcomes of AI-based decision-making or the dysfunctionality of AI systems and models.

b) Corrective and non-corrective AI contestability

Against this backdrop, we argue that the provisions of the AI Act read in the broader context of international human rights law also play a role with regard to non-corrective forms of AI contestation, which integrate a second distinct perspective on AI contestability particularly useful to address challenges stemming from the use of AI for law enforcement purposes. By definition, non-corrective AI contestability does not seek to enhance the performance, underlying logic or objectives of an AI system, model or tool. Non-corrective AI contestability can take different forms, including spontaneous contestation as

¹¹⁵Ibid.

¹¹⁶Palmiotto (fn. 6), pp. 790 ff.

¹¹⁷Simon Hirsbrunner/Steven Kleemann/Milan Tahraoui, Contestation in artificial intelligence as a practice: from a system-centered perspective of contestability towards normative contextualization, situative critique and organizational culture, in: *Frontiers in Communication* 10 (2025), pp. 1-15 (8).

¹¹⁸Claudia Aragau, Error, in: Mareile Kaufmann/Heidi Mork Lomell (ed.), *De Gruyter Handbook of Digital Criminology*, 2025, pp. 215-221 (216, 219).

¹¹⁹Gianclaudio Malgieri/Frank Pasquale, Licensing high-risk artificial intelligence: Toward *ex ante* justification for a disruptive technology, in: *Computer Law & Security Review* 52 (2024), pp. 1-18.

social practice or 'techno-resistance',¹²⁰ and can pursue various objectives, such as preventing an AI development or AI deployment project altogether, or targeting more specific deployment contexts or modalities of use.

In the context of the predominant AI regulatory approaches based on risks, this perspective on non-corrective AI contestability is particularly useful and relevant for analysing social practices in light of the relationship between legal imaginaries, traditional modes of law-making and innovation, especially given the fact that most of AI regulatory frameworks focus on the concepts of AI trustworthiness and acceptability. According to the 2019 AI High-Level Expert Group, trustworthiness can be defined as lawful, ethical, and robust AI (technically and socially speaking) throughout the AI lifecycle.¹²¹ Among the requirements suggested by these experts, which influenced the 2021 EU Commission Proposal for an AI Regulation,¹²² one refers to accountability and includes the criteria of "auditability, minimisation and reporting of negative impact, trade-offs, and redress".¹²³ The emphasis placed on trust and trustworthiness in various international and regional AI regulatory frame-

works translates into a general objective of inducing people to trust AI, innovation and to use this technology, thereby unlocking its economic and societal potential. However, one fundamental issue is that trust cannot be commanded; it requires fulfilment of the necessary conditions. It has been argued that constant appeals to the concept of trustworthiness can lead to confusion between this concept and acceptability.¹²⁴ This problem is exacerbated by the fact that trust is meant as an expert domain under the AI Act and other regulatory frameworks. AI contestability as a social practice is therefore relegated to the background or ignored entirely. Yet one might surmise that contestability is necessary in a democratic context, and inevitable in a non-democratic context. Overreliance on trust and the risks stemming from AI opacity create a basic need for institutions, mechanisms, norms, and cultures that enable effective contestation. Contestability is especially crucial for prohibited AI practices and high-risk AI systems used for security purposes. This is arguably due to the coercive nature of AI tools developed in this context and the power imbalances they generate when deployed, given the involvement of State authorities and the often dominant private corporations.¹²⁵ A major advantage of the concept of AI contestability is that it incorporates both 'institutional' and 'formal rules-based reactions' to AI development and deployment, as well as 'spontaneous, informal and cultural interactions and responses'.

¹²⁰See for instance, *Marie Petersmann*, Refusing Algorithmic Recognition, in: *European Journal of International Law* 35 (2024), pp. 979-989 (986 f.).

¹²¹High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI* of 8 April 2019, available at: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> (last visited: 12 December 2025), p. 5.

¹²²European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, COM(2021) 206 final of 21 April 2021.

¹²³High-Level Expert Group on Artificial Intelligence (fn. 121), p. 14.

¹²⁴*Johann Laux/Sandra Wachter/Brent Mittelstadt*, Trustworthy artificial intelligence and the European Union AI Act: On the conflation of trustworthiness and acceptability of risk, in: *Regulation and Governance* 18 (2024), pp. 3-32 (27).

¹²⁵*Ibid.*, see also Art. 7 para. 2 lit. h) and Recitals 59-60 AI Act.

Unlike corrective forms of AI contestability, the concept of non-corrective contestability is not limited to situations where something might go wrong. This enables criticism of the very foundations of an AI development or deployment project to be taken into account, while offering a better focus on systemic risks relating to sensitive intended AI use cases and how they are perceived by legal imaginaries and traditional modes of law-making. Non-corrective AI contestability can pursue various objectives, such as preventing an AI development or deployment project altogether, or targeting more specific deployment contexts or modalities of use. Non-corrective AI contestability is far less anchored in an *ex post* logic, thus offering more critical space for *ex ante* forms of friction, opposition, resistance or refusal.¹²⁶ For instance, a key discussion about AI contestability is whether it is possible to effectively contest the very idea of developing or acquiring an AI system or model. A notable example of this is the contestability of AI procurement procedures, given that one of the initial stages of procurement entails identifying the actual needs behind the procurement of such systems or models, along with their requirements.¹²⁷ This is particularly difficult in fields of AI development and deployment in the private and public security sectors, as third parties – and even less the broader public – are rarely involved.

First, there are limits to corrective AI contestability, such as empirical or legal ones. For example, Art. 86 AI Act establishes

¹²⁶Petersmann (fn. 120), pp. 983 f.

¹²⁷Digital Regulation Cooperation Forum (DRC), Transparency in the procurement of algorithmic systems: Findings from our workshops, 2023, available at: <https://www.drcf.org.uk/siteassets/drcf/pdf-files/transparency-procurement-algorithmic-systems.pdf?v=381844> (last visited 23 August 2025).

a right to an explanation of individual AI-based decision-making. However, its contestability potential is significantly limited by the fact that this provision only entitles affected persons to obtain clear and meaningful explanations from the deployer regarding the role of the AI system in the decision-making process and the main elements of the decision made. This right to explanation provides a significant legal mechanism that could potentially improve understanding of whether an AI-based decision took place, but also on what basis and according to which rules. There can be, therefore, a potential for contestation. This is, however, only the case indirectly and *ex post facto*, and more generally, without allowing the questioning of the foundations of those AI-based decisions. In fact, there is also a right to lodge a complaint on the basis of Art. 85 AI Act. That said, the AI Act does not actually grant a right to directly contest the development or deployment of an AI system. Rather, this Regulation establishes a right to a mediated formal, legal and institutional contestation, i.e., through an expert or official representation (e.g. oversight bodies). What is then particularly problematic is the fact that the AI Act does not impose on market surveillance authorities either to report on how they handle complaints, or to provide the possibility to ‘appeal’ their decisions. Indeed, over the course of the negotiations, the European Parliament wished to introduce a “right to an effective judicial remedy against a national supervisory authority”,¹²⁸ in addition to the right to lodge a complaint. However, the final version of the AI Act

¹²⁸European Parliament, Amendments adopted by the European Parliament on 14 June 2023 on the proposal of the European Parliament and the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), (Ordinary legislative procedure: first reading), OJ C/2024/506 of 23 January 2024, Amendment 629.

has retained a system of remedies that confers more responsibility on the surveillance market authorities, while reducing possibilities for complainants to act upon these authorities' decisions.¹²⁹

Secondly, there are also empirical limits observable in how AI-based technologies transform security politics, affect human rights protection and more generally reshape political and social interactions. For example, the introduction of AI technologies has led to a new surveillance logic in the areas of migration, asylum and border management control, which the AI Act generally classifies as high-risk (Recital 60 AI Act). In this context, there are significant challenges to AI contestability, as the logic of AI-based border surveillance systems aims to discover 'unknown unknowns' not solely on the basis of pre-defined risk concepts and categories, but also based on a "*dispositif* of pre-emptive security or speculative suspicion"¹³⁰ or "inferred attributes"¹³¹ that are used to generate fluid categories for sorting persons under surveillance. Similarly, *Rinaldi* and *Teo* argue that "the deployment of AI-driven border and migration management may be challenging the idea of individual empowerment which lies at the core of the human rights protection framework",¹³² through datafication,¹³³ inference and construc-

tion¹³⁴ as well as algorithmic groupings.¹³⁵ In light of this, how can an AI-based socio-technical system that is developed or deployed for public security purposes—and whose underlying logic is difficult to understand—be meaningfully contested? When the system's secrecy is protected either as a matter of public security policy¹³⁶ or as a trade secret under public-private partnership agreements, the inherent opacity of such a socio-technical system can sharply shrink the room for corrective interventions.

IV. Conclusion

The EU AI Act marks a pivotal step towards governing AI systems, yet its application to law enforcement contexts remains fraught with ambiguity and competing priorities. While the AI Act's risk-based architecture offers a structured baseline, the introduction of exemptions ("backdoors") for security agencies dilutes that clarity. These carve-outs allow high-risk technologies, including biometric surveillance and predictive policing tools, to bypass safeguards that would otherwise apply to private actors, thereby widening the gap between the Act's stated commitment to fundamental rights protection and its practical enforcement.

Our analysis shows that the current accountability regime — comprising provider-

¹²⁹See in this sense, European Center for Not-for-Profit Law (fn. 76), p. 39; *Griff Ferris/Sofia Lyall*, *New Technology, Old Injustice. Data-driven discrimination and criminalisation in police and prisons in Europe*, Statewatch of June 2025, available at: <https://www.statewatch.org/news/2025/june/police-racism-and-criminalisation-a-cross-europe-increasingly-fuelled-by-digital-prediction-and-profiling-systems/> (last visited 21 August 2025), p. 20.

¹³⁰*Sullivan/Van Den Meerssche* (fn. 21).

¹³¹*Amoore* (fn. 21).

¹³²*Rinaldi/Teo* (fn. 15), p. 78.

¹³³*Ibid.*

¹³⁴*Ibid.*, p. 79.

¹³⁵*Ibid.*, p. 80.

¹³⁶Statewatch, EU's secretive "security AI" plans need critical, democratic scrutiny says new report, Statewatch of 29 April 2025, available at: <https://www.statewatch.org/news/2025/april/eu-s-secretive-security-ai-plans-need-critical-democratic-scrutiny-says-new-report/> (last visited 23 August 2025).

centric risk-management obligations, limited post-market monitoring, and a nascent fundamental rights impact assessment — does not fully empower affected individuals to contest AI-driven decisions. The reliance on providers to embed safeguards, coupled with the limited scope of supervisory authorities, creates a systemic asymmetry: law enforcement bodies can deploy powerful AI tools while citizens face procedural hurdles to obtain redress.

Meaningful protection of fundamental rights will depend on several interrelated reforms. First, exemptions should be tightened, and ‘backdoor’ provisions need to be narrowed or eliminated so that law enforcement applications are subject to the high-risk requirements intended for them. Second, contestability mechanisms must be strengthened by expanding the scope of the fundamental rights impact assessment, guaranteeing transparent documentation, and providing individuals with enforceable rights of review and remediation. Third, the responsibilities of providers and deployers should be clarified, with precise duties assigned to each participant in the AI supply chain, especially where deployers such as police forces possess superior contextual knowledge of the risks. Finally, a rights-based overlay ought to be embedded within the risk-based framework, ensuring that fundamental liberties are treated as non-negotiable rather than merely a variable in a cost-benefit analysis. Only by reconciling the Act’s technical risk calculus with a robust, rights-centred accountability architecture can the EU ensure that AI enhances public safety without eroding the democratic values it seeks to protect.

Vitae

Steven Kleemann is a doctoral researcher at the Faculty of Law at the University of Potsdam as well as a researcher and policy advisor on digitalisation, AI & human rights at the German Institute for Human Rights. At the time of writing this article, he was a researcher at the Berlin Institute for Safety and Security Research (FÖPS Berlin), working on a project concerning legal aspects of trustworthy AI for police applications. His research focuses on international law, human rights, AI, and security law.

Milan Tahraoui is a doctoral researcher associated with the Centre Marc Bloch, as well as a Ph.D. candidate at both the Paris 1 Pantheon-Sorbonne University and the Free University of Berlin. At the time of writing this article, he was a researcher at the Berlin Institute for Safety and Security Research (FÖPS Berlin), working on a project examining the issues surrounding the use of so-called trustworthy AI applications in law enforcement. His research focuses on international and European law, human rights, digital surveillance, AI and security law.

The Impact of Artificial Intelligence on the Work of Human Rights Defenders

Manuel Brunner¹

¹Prof. Dr., LL.M., HSPV NRW

Contents

- I. Introduction
- II. Definition of and Legal Framework for Human Rights Defenders
- III. AI as a Tool for Human Rights Defenders
- IV. AI as a Threat to Human Rights Defenders
- V. International AI Legislation and Human Rights Defenders
- VI. Conclusion
Vita

Abstract

Artificial Intelligence (AI) is one of the most important technologies of the present day. It is already influencing the day-to-day life of many people and is shaping the decision-making of enterprises, governments, humanitarian organisations and other actors. This article discusses the potentials of the use of AI by and against human rights defenders. On the one hand, AI can offer immense benefits for such individuals or groups with its timesaving, simplifying and analytical capabilities. On the other hand, AI can be used to track and attack human rights defenders and their activities. It offers an additional way, especially for governments, to curtail the rights of human rights defenders and to hinder them in their work.

Keywords

Artificial Intelligence, Human Rights Defenders, Technology, Data Analysis, Human Rights Campaigning, Human Rights Violations, AI Legislation, Visualization

Citation:

Manuel Brunner, The Impact of Artificial Intelligence on the Work of Human Rights Defenders, in: MRM 30 (2025) 2, pp. 144–158.
<https://doi.org/10.60935/mrm2025.30.2.32>.

Received: 2025-11-24

Accepted: 2026-01-26

Published: 2026-02-17

Permissions:

The copyright remains with the authors.
Copyright year 2026.

Unless otherwise indicated, this work is licensed under a [Creative Commons License Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/). This does not apply to quoted content and works based on other permissions.

I. Introduction*

Current discussions about artificial intelligence (AI) often revolve around questions of the impact of this capability of computational systems on the future of work or on the dangers AI poses for the enjoyment of human rights.¹ However, a rarely discussed point in those debates is the impact that AI has on the work of those individuals or organisations who endeavour to improve the human rights situation for others, namely the impact on human rights defenders. On the one hand, AI can be a powerful tool for human rights defenders supporting them in fields like data analysis or campaigning or can be used to provide services for people who face situations which are problematic in terms of the enjoyment of human rights. Moreover, AI might be used to visualize human rights violations and AI-driven application can serve as warning tools when human rights defenders are under threat. On the other hand, states might use AI to persecute and

curtail the rights of those actors. While the field of AI and its risks to human rights have already been discussed by some Special Rapporteurs of the United Nations Human Rights Council regarding their respective mandates, an articulation by the Special Rapporteur on the situation of human rights defenders is yet missing.² For the purpose of the following, the term “artificial intelligence” is understood along the lines of the definition of AI systems in Article 2 of the Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (Framework Convention).³ Therefore, AI “means a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations or decisions that may influence physical or virtual environments. Different artificial intelligence systems vary in their levels of autonomy and adaptiveness after deployment.”

This article discusses the various aspects of the impact of AI on the work of human rights defenders as outlined above. To add context to the following descriptions and analyses the explanations start with a definition of the term “human rights defender” and an introduction to the legal framework under which those actors operate. In the following section AI is discussed as a useful tool in the hands of human rights

* This paper was presented at the 30th Anniversary Conference of the Human Rights Centre of the University of Potsdam “Human Rights and Artificial Intelligence Addressing challenges, enabling rights”, 7–8th November 2024 in Potsdam, Germany.

¹ See for instance *Thomas H. Davenport*, *The AI Advantage: How to Put the Artificial Intelligence Revolution to Work*, 2018; International Monetary Fund (ed.), *Gen-AI: Artificial Intelligence and the Future of Work: Staff Discussion Note*, 2024; *Jean-Philippe Deranty/Thomas Corbin*, *Artificial intelligence and work: a critical review of recent research from the social sciences*, in: *AI & Society* 39 (2024), pp. 675–691; *Steven Livingston/Mathias Risse*, *The Future Impact of Artificial Intelligence on Humans and Human Rights*, in: *Ethics & International Affairs* 33 (2019), pp. 141–158; *Rowena Rodrigues*, *Legal and human rights issues of AI: Gaps, challenges and vulnerabilities*, in: *Journal of Responsible Technology* 4 (2020), 100005; *Onur Bakiner*, *The promises and challenges of addressing artificial intelligence with human rights*, *Big Data & Society* 10 (2023).

² Reports about AI in the fields of their respective mandates have for instance been: UNHRC, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression of 29 August 2018, UN Doc. A/73/348; UNHRC, Report of the Special Rapporteur on the right to education of 16 October 2024, UN Doc. A/79/520; UNHRC, Report of the Special Rapporteur in the field of cultural rights of 30 July 2025, UN Doc. A/80/278.

³ Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law of 5 September 2024, CETS No. 225.

defenders. Several potential and actual applications are presented in that regard. Topics like AI-driven data analysis, exemplified by the work of non-governmental human rights organisations, and the creation of a warning tool for human rights defenders are discussed. The next section of the article explores the use of AI against human rights defenders. Examples from known practices of authoritarian governments are used to illustrate the dangers of AI-driven technologies for the work of human rights defenders. Thereafter, the challenges of current international legislation on AI regarding human rights defenders are discussed. The article ends with a conclusion.

II. Definition of and Legal Framework for Human Rights Defenders

The term “human rights defender” has gained prominence in human rights work, politics, international relations and legal studies for over a quarter century now.⁴ Before that, individuals or groups who would today be labelled as human rights defenders were more regularly called human rights activists, professionals, workers or monitors.⁵ While those terms have

not vanished, the new term is arguably seen and used as the more common one today. The origin of the described change in the usage of terms lies in the year 1998. After 14 years of negotiations, the General Assembly of the United Nations adopted the Declaration on the Right and Responsibility of Individuals, Groups and Organs of Society to Promote and Protect Universally Recognized Human Rights and Fundamental Freedoms on 9 December of that year.⁶ This official title of the instrument is often shortened to “Declaration on human rights defenders”.⁷ The shortening serves practical purposes as it clearly underlines who benefits from its regulations. The term “human rights defender” is not defined in the declaration itself. However, in the practice of the United Nations human rights defenders are considered “people who, individually or with others, act to promote or protect human rights in a peaceful manner.”⁸

The Declaration on human rights defenders is an international instrument of the right to defend human rights. It reaffirms rights that are of utmost importance for the defence of the rights of others.⁹ Those rights include the freedom of association, the freedom to peaceful assembly, the freedom of opinion and expression, the right to gain access to information, the right to provide legal aid and the right to develop and discuss new ideas in the area of human rights.¹⁰ The implementation of the Declaration on human rights de-

⁴ See for instance the contributions in *Alice M. Nah (ed.)*, *Protecting Human Rights Defenders at Risk*, 2020; *Benjamin Beuerle*, *Zur Umsetzung der „Erklärung zu den Menschenrechtsverteidigern“ fünf Jahre nach ihrer Verabschiedung – eine Bestandsaufnahme*, in: MRM 9 (2004), pp. 47–52; *Norman Weiß*, *Schutz von Menschenrechtsverteidigern – neuere Entwicklungen*, in: MRM 21 (2016), pp. 29–44.

⁵ OHCHR, *About human rights defenders*, available at: <https://www.ohchr.org/en/special-procedures/sr-human-rights-defenders/about-human-rights-defenders> (last visited 15 November 2025).

⁶ UN Doc. A/RES/53/144.

⁷ For example UN, *Commentary to the Declaration on the Right and Responsibility of Individuals, Groups and Organs of Society to Promote and Protect Universally Recognized Human Rights and Fundamental Freedoms of July 2011*, p. 5.

⁸ OHCHR (fn. 5).

⁹ UN (fn. 9).

¹⁰ *Ibid.*

fenders is a precondition for the creation of an environment that enables human rights defenders to carry out their work.¹¹ As the Declaration on human rights defenders is a document that was adopted by the General Assembly of the United Nations it is not legally binding and is therefore a soft law document.¹² However, as individuals or groups, human rights defenders also benefit from the rights enshrined in treaties and other international instruments in the field of human rights such as the International Covenant on Civil and Political Rights¹³ or the Universal Declaration of Human Rights¹⁴ and from fundamental rights in national constitutional law. Furthermore, some states have introduced legislation or policies for the benefit of human rights defenders.¹⁵

¹¹ Ibid.; UN Doc. E/CN.4/2006/95, para. 50.

¹² See for a detailed analysis *Stephen M. Schwebel*, The Effect of Resolutions of the U.N. General Assembly on Customary International Law, in: Proceedings of the Annual Meeting of the American Society of International Law 73 (1979), pp. 301–309; *Ludovic Hennebel/Hélène Tigroudja*, International Human Rights Law: A Treatise, 2025, pp. 92 ff.

¹³ International Covenant on Civil and Political Rights of 16 December 1966, UNTS vol. 999, p. 171

¹⁴ Universal Declaration of Human Rights of 10 December 1948, UN Doc. A/RES/217 A (III).

¹⁵ See for instance the following legislation: United Mexican States, Ley para la protección de personas defensoras de derechos humanos y periodistas of 25 June 2012, Diario Oficial de la Federación ID 270; Republic of Honduras, Ley de la protección para las y los defensores de derechos humanos, periodistas, comunicadores sociales y operadores de justicia of 15 May 2015, Diario Oficial de la República de Honduras, Número 33, 730, A. 1–21 or Republic of Niger, Loi fixant les droits et devoirs des défenseurs des droit de l'homme au Niger of 20 June 2022, Journal officiel de la République du Niger, Loi N° 2022-27; see also International Service of Human Rights, Model Law for the Recognition and Protection of Human Rights Defenders of 21 June 2016, available at: <https://ishr.ch/defender-s-toolbox/resources/model-law/> (last visited 15

Guidelines and policies for the protection of human rights defenders have also been issued by several international organisations and institutions.¹⁶

III. AI as a Tool for Human Rights Defenders

As stated above, AI can serve various needs of human rights defenders in carrying out their work. AI tools and applications may be used for analysing data, preparing and carrying out human rights campaigns and providing services to individuals who find themselves in situations which are critical

November 2025); and the following guidelines: the Guidelines for Irish Embassies and Missions on Human Rights Defenders of 2010, available at: https://www.humanrights.ch/cms/upload/pdf/150415_irish_hrd_guidelines_en.pdf (last visited 15 November 2025); the Swiss Guidelines on the Protection of Human Rights Defenders of 2013, available at: <https://www.eda.admin.ch/eda/en/fdfa/fdfa/publikationen/alle-publikationen.html/content/publikationen/en/eda/menschenrechte-humanitaeres-migration/Leitlinien-zum-Schutz-von-HRD/> (last visited 15 November 2025) or the Guidelines on Supporting Human Rights Defenders of Canada of 2017, available at: https://www.international.gc.ca/world-monde/issue_s_development-enjeux_developpement/human_rights-droits_homme/rights_defenders_guide_d_edefenseurs_droits.aspx?lang=eng (last visited 15 November 2025).

¹⁶ See for example Ensuring Protection - European Union Guidelines on Human Rights Defenders of 2008, available at: https://www.eeas.europa.eu/sites/default/files/eu_guidelines_hrd_en.pdf (last visited 15 November 2025); Guidelines on the Protection of Human Rights Defenders of the Organization for Security and Co-operation in Europe of 10 June 2014, available at: <https://odhr.osce.org/odhr/guidelines-on-the-protection-of-human-rights-defenders> (last visited 15 November 2025); Committee of Ministers of the Council of Europe, Recommendation to member States on the need to strengthen the protection and promotion of civil society space in Europe, CoE Doc. CM/Rec(2018) 11 of 28 November 2018.

concerning the enjoyment of human rights of those persons. Visualizing human rights violations and usage as a warning tool complements the usage of AI by and for human rights defenders. In the following section of this article, some applications of AI technology in the work of human rights defenders are discussed. The list of applications presented here is, of course, illustrative, not exhaustive.

1. Data Analysis

Human rights defenders often undertake to collect and analyse data of events that might be human rights violations or violations of international humanitarian law. Data about potential human rights violations might come from a myriad of diverse sources. Such sources may include interviews, photographs, sound files, video footage, written accounts, maps or satellite imagery. Once the data is analysed, it might be used to produce reports, as evidence in court procedures, in educational materials or in other ways. Especially in situations in which a great number of human rights violations take place, such as in armed conflicts or within the context of authoritarian governments, many potential pieces of evidence might be available. Analysing such huge quantity of data can be done by humans exclusively. However, the process can be time-consuming and requires many human recourses. Therefore, non-governmental organisations in the field of human rights make use of AI-driven technologies to support their experts and researchers to analyse data in these kinds of situations.

An example for such data analysis can be found in the practice of the non-governmental human rights organisation Amnesty In-

ternational.¹⁷ The organisation used AI within the context of the human rights situation in Sudan, more precisely in the country's western region of Darfur.¹⁸ The armed conflict in Darfur between the Sudanese government and opposition groups in the region has started in 2003. A core element of the situation is high level violence against civilians. The atrocities committed in Darfur include killings, torture, enforced disappearances, rape and other forms of sexualized violence, pillage, forced displacement and the destruction of villages.¹⁹ In 2016, Amnesty International documented a new wave of violence which included attacks on and the destruction of many villages in Darfur.²⁰ To understand the scale of the destruction, Amnesty International launched the projects "Decode Darfur" and "Decode the Difference". The organisation called upon volunteers to scan satellite images of Darfur and to identify destroyed villages in the process. The task set for those volunteers was to study the available images and to determine if villages had been

¹⁷ *Anne Dulka*, The Use of Artificial Intelligence in International Human Rights Law, in: *Stanford Technology Law Review* 26 (2023), pp. 316-366 (330).

¹⁸ *Ibid.*, p. 331.

¹⁹ See for instance *Roberto Belloni*, The Tragedy of Darfur and the Limits of the 'Responsibility to Protect', in: *Ethnopolitics* 5 (2006), pp. 327-346; *Alex de Waal*, Darfur and the failure of the responsibility to protect, in: *International Affairs* 83 (2007), pp. 1039-1054; *Joyce Apsel*, The Complexity of Destruction in Darfur: Historical Processes and Regional Dynamics, in: *Human Rights Review* 10 (2009), pp. 239-259.

²⁰ *Milena Marin/Freddie Kalaitzis/Bufy Price*, Using artificial intelligence to scale up human rights research: a case study on Darfur, Amnesty International Evidence Lab of 6 July 2020, available at: <https://citizenevidence.org/2020/07/06/using-artificial-intelligence-to-scale-up-human-rights-research-a-case-study-on-darfur/> (last visited 15 November 2025). The following paragraphs are based on that report.

destroyed between 2014 and 2016. The project was joined by 28,600 individuals from 147 countries. They generated 13 million annotations covering 2.6 million satellite images of approximately 100x100 meters each. With this method an area of 300,000 square kilometres was covered. However, this was not enough to cover the entire 493,180 square kilometres of the region. To scale up the work, Amnesty International made use of AI.

Before sharing the collected data with partners and engineers, Amnesty International developed a risk assessment model. In this model four criteria for sharing the data were identified, namely sensitivity and graphic nature of the data, transparency, data validation and risks of malicious attacks and misuse of data. Amnesty International then worked with experts on machine learning from the University College London in the United Kingdom during the next steps of the project. The large dataset of annotated chips of satellite imagery from the two prior projects on Darfur was used to train a machine learning model to automatically map the vast desert areas of the region. The model was able to identify “human presence”, as well as “destroyed” or “mixed” (i.e., partially destroyed) villages. This was done for the whole of Darfur. The results were embedded in a web-based mapping application. This tool allowed the researchers of Amnesty International to refine the annotations and visualize at scale the patterns of habitation and the destruction. However, the work still faced considerable limitations due to a small team with limited resources, a reliance on a mosaic of archival images that were stitched together, and the fact that some tiles dated back more than ten years. To overcome those challenges, Amnesty International entered into a relationship with the then-existing Canadian AI company Element AI.

Through the collaboration of Element AI with the Satellite Applications Catapults’ ML Use Case Programme, Amnesty International was able to access high resolution commercial data. These data allowed the Element AI team to assess the viability of training deep learning models for detecting destroyed villages in Darfur. With a dataset of images and crowdsourced annotations, indicating whether a human settlement was visible in an image, and if so, whether the settlement was destroyed, it was the aim of Amnesty International to learn the task of mapping an image to the correct annotation. This meant that the team of Amnesty International wanted to accurately predict whether a human annotator would mark that image as a destroyed village. After more fine-tuning, the organisation concluded that the best used system deployed in the general Darfur area would successfully identify 82% of all destruction cases (and miss 18% of them), while only 15% of the destruction “alerts” would be false positives. These metrics indicate that the deployed model could thereby save significant time and operational costs for intelligence research.

After the exercise, Amnesty International felt that large scale human rights research and continuous monitoring of conflicts aided by AI were within reach. However, numerous challenges occur when such technology is used for those tasks. According to Amnesty International, four main lessons were learned during the exercise. First, human rights organisations need to carefully evaluate the risks of such research against potential benefits. If effective migration strategies cannot be put in place, human rights organisations should be comfortable dropping such projects altogether. Secondly, training data for human rights AI algorithms is often not easily available, and data collection processes are not straightforward. Amnesty Inter-

national benefited from its already-large base of volunteers to generate such data. In the future, the organisation wants to use a combination of volunteer-driven analysis and analysis by AI. The idea for such work is that the AI can supply functionality in scale, whereas the volunteers provide depth to the analysis. Thirdly, Amnesty International pointed out that projects like the one on Darfur would require a large constellation of partnerships and pro-bono or heavily subsidized collaborations. Fourthly, the accuracy of algorithms should be carefully considered. Human rights researchers would need technical literacy to understand and interpret the data and to make crucial design choices such as calibration of accuracy metrics such as precision and recall.

Another example is the work of Human Rights Watch regarding Myanmar.²¹ The human rights non-governmental organisation also partnered with Element AI to monitor ethnic violence in the country.²² This was done against the background of violence against the Rohingya population in the state of Rakhine.²³ Human Rights Watch worked with the company to design a machine learning tool that could use satellite imagery and remote sensing thermal data to spot, track, and catalogue violations of human rights against the Rohingya.²⁴

Human Rights Watch used thermal data to monitor the violence with a focus on

attacks on settlements.²⁵ Indicators such as smoke plums captured by environmental satellites were used in the procedure. The researchers were then able to make use of AI to combine this data with aerial images to identify spots where violence had taken place. Furthermore, AI was used to combine the data with other publicly available information, like videos and photos which were posted on social media. Thereby, Human Rights Watch was able to pinpoint where burnings of villages were conducted. This then allowed it to corroborate the testimony of individuals who were victims of human rights violations and to identify perpetrators. Human Rights Watch was able to identify at least 214 villages that were nearly totally destroyed as part of the ethnic cleansing campaign against the Rohingya by the Burmese military.

2. Human Rights Campaigning

Another important aspect of the work of human rights defenders is campaigning. Campaigning can for instance concern a particular situation in which human rights are in danger at a large scale, the fate of an individual who is in grave danger of having his or her human rights violated or general information about human rights. A campaign can bind many resources which cannot be used for other tasks while the campaign needs to be prepared and executed. AI can be used during the preparation and the running phase of a human rights campaign to support the work of human rights defenders in several ways.

²¹ *Dulka* (fn. 17), p. 337.

²² *Ibid.*

²³ See for instance *Anthony Ware/Costas Laoutides*, Myanmar's 'Rohingya' Conflict, 2018; *Ken MacLean*, The Rohingya Crisis and the Practices of Erasure, in: *Journal of Genocide Research* 21 (2019), pp. 83-95.

²⁴ *Dulka* (fn. 17), pp. 337, 338.

²⁵ Human Rights Watch, Burma: Satellite Imagery Shows Mass Destruction - 214 Villages Almost Totally Destroyed in Rakhine State of 19 September 2017, available at: <https://www.hrw.org/news/2017/09/19/burma-satellite-imagery-shows-mass-destruction> (last visited 15 November 2025). The following paragraph is based on that report.

AI might be used as a tool for strategic communication.²⁶ In that regard AI tools can help human rights defenders with their ability to produce ideas within an instant and thereby reducing the time and resources needed to mobilize support and drive social change. Furthermore, it is a typical feature of human rights campaigns that they call to action for individuals. The individuals who are targeted with such a campaign might hold different values, different views on political, economic or social questions which are relevant for the specific campaign and have made different experiences in life. AI can assist human rights defenders in explaining the issues at stake in the campaign, which might involve complex points, and in convincing individuals of the goals of the effort. Such a complex point might be the relationship between climate change and the enjoyment of human rights. If human rights defenders, for instance, endeavour to convince individuals who are sceptical of the existence of climate change, they can harness the power of AI prompt engineering to better tailor their respective messages in a way that resonates more with the targeted audience.

AI technology can also play a role for human rights defenders in administrative and logistical tasks during campaigning. Through the use of such technologies, human rights defenders can allocate more time and energy to strategic and tactical planning. Repetitive tasks that may be fulfilled by AI tools or in which such tools might be helpful to include generating responses to e-mails, summarizing long

texts and reports, project management or creating presentations. Moreover, AI-driven audio transcription technologies can help to make produced content more easily accessible in a variety of formats. In addition, AI-driven chatbots can be used as a first point of contact. The chatbots can engage with supporters and stakeholders and address frequently asked questions or answer initial inquiries.

Furthermore, AI can be used by human rights defenders to create political satire. The combination of humour and political protest has long been a catalyst for social and political change.²⁷ Furthermore, the use of humour, satire or parody can lead critical discourse about the absurdity and consequences of policy decisions.²⁸ Human rights defenders might use the capabilities of AI to create satirical songs, pictures, videos, small games or internet memes.²⁹ For instance, AI pictures might be generated that show political leaders in mock scenarios where they are presented as low-wage workers or refugees to visualize the impact of their political decisions or to unmask their rhetoric.³⁰ One example which was used by a human rights group concerned British politician Suella Braverman of the Conservative and Unity

²⁶ *Melissa McNeilly*, *The Human Rights Opportunities of Artificial Intelligence (AI), New Tactics in Human Rights* of 28 August 2023, available at: <https://www.newtactics.org/perspectives/human-rights-opportunities-artificial-intelligence-ai/> (last visited 15 November 2025). The following paragraphs are based on that article.

²⁷ *Ibid.*; see also on this topic *Atyaf Fakhri Hassan Hussein*, *The Use of Satire in Addressing Corruption Issues and the Benefit of Artificial Intelligence: An Analytical Study in the Program "Al-Rashid and the People"*, in: Fausto Pedro García Márquez/Alaa Ali Hameed/Akhtar Jamil (ed.), *Pattern Recognition and Artificial Intelligence: Selected papers from the 6th Mediterranean Conference on Pattern Recognition and Artificial Intelligence (MedPRAI24)*, 2026, pp. 359–376; *Hang Lu/Shupeiyuan*, *"I know it's a deepfake": the role of AI disclaimers and comprehension in the processing of deepfake parodies*, in: *Journal of Communication* 74 (2024), pp. 359–373.

²⁸ *McNeilly* (fn. 26).

²⁹ *Ibid.*

³⁰ *Ibid.*

Party. She has been a Member of Parliament and served as Secretary of State for the Home Department in the cabinets of Prime Ministers Mary Elizabeth Truss and Rishi Sunak.³¹ Braverman is known for her very critical stance on immigration, refugee issues and multiculturalism.³² The human rights group has let an AI tool create a picture of a woman who looks like Braverman.³³ This woman has a toddler in her arms and wears a life jacket. The picture shows her in front of a body of water with a coastline in the background. Thereby, it is intended that the woman in the picture looks like a refugee. The AI-generated image was meant to criticize Braverman's harsh immigration policies. Using such pictures in human rights campaigns might be seen as controversial; however, it can succeed in capturing public attention.³⁴

3. Visualizing Human Rights Violations

AI-generated images or films cannot only be used by human rights defenders as a tool for tasks in campaigns as outlined in the previous paragraph. Such media files might prove also highly useful to visualize human rights violations. The ever-growing capabilities of AI applications in the sphere of generating imagery, videos and sound files can play a pivotal role in this regard.

Human rights violations often take place away from the public eye. Authoritarian governments, non-state armed groups or other actors might try to hide their atrocities to avoid prosecution by national or international institutions of criminal justice or condemnation by the international community or public outrage and thereby, to steer clear of losing support for their respective causes. An example for such behaviour includes the usage of the infamous Colonia Dignidad, a settlement of the likewise named Christian cult by German immigrants near the city of Parral in Chile. The Chilean government under dictator General Augusto Pinochet used the area as a centre for internment, torture and murder of dissidents during the 1970s.³⁵ Another example was the secret prison complex El Vesubio in the metropolitan area of Buenos Aires. The military dictatorship of Argentina used the site to torture and murder political rivals from 1976 to 1978.³⁶ Human rights violations may also occur in the context of a general policy of public secrecy in and isolation of a country. This is, for instance, the case for North Korea where many human rights violations were made public by individuals who have managed to flee the country.³⁷ A

³¹ Government of the United Kingdom, The Rt Hon Suella Braverman KC MP, available at: <https://www.gov.uk/government/people/braverman> (last visited 15 November 2025).

³² See Philip Hubbard, Suella Braverman's talk of a refugee 'invasion' is a dangerous political gambit gone wrong, King's College London of 3 November 2022, available at: <https://www.kcl.ac.uk/suella-bravermans-talk-of-a-refugee-invasion-is-a-dangerous-political-gambit-gone-wrong> (last visited 15 November 2025).

³³ McNeilly (fn. 26).

³⁴ Ibid.

³⁵ See for a detailed analysis Evelyn Hevia Jordán, Colonia Dignidad: Lights and Shadows in the Recognition of the Victims, in: Elizabeth Lira/Marcela Cornejo/Germán Morales (ed.), Human Rights Violations in Latin America: Reparation and Rehabilitation, 2022, pp. 223–236; Caroline Moine, Denouncing or Supporting the Chilean Dictatorship in West Germany? Local Associations of Solidarity and Their Transnational Networks Since the 1970s, in: Global Society 33 (2019), pp. 332–347.

³⁶ See in greater detail Gonzalo Conte, A topography of memory: Reconstructing the architectures of terror in the Argentine dictatorship, in: Memory Studies 8 (2015), pp. 86–101.

³⁷ See for more detailed accounts the UNHRC, Report of the Commission of Inquiry on Human Rights in the Democratic People's Republic of Ko-

context in which human rights violations also take place away from the public eye is the situation of refugees who try to reach the territory of member states of the European Union via the Mediterranean Sea. Human rights violations against refugees on board of ships, boats and skiffs by the Libyan Coast Guard have been reported on numerous occasions.³⁸ The list of such human rights violations could be extended with ease.

Reporting on human rights violations that take place away from the public eye often relies on the testimony of victims or witnesses alone or to a great degree. While such accounts can be powerful in themselves, a visualization of the told stories has the potential to add additional weight to the reports. Human rights defenders can use AI-based tools to reconstruct the interior of a prison camp, the site of a massacre or the conditions on board of coast guard vessels in images or films. The so created media files can then for instance be used to raise public awareness for a specific situation, in educational events about human rights issues or in presentations at hearings of parliamentary committees. Usually, the task of visualizing human rights violations was carried out by draftsman who manually, or by using graphic software, produced images based on the accounts of the victims or witnesses. However, using AI for such a role offers the advantage that images can be created more quickly and in

greater numbers, while reducing costs for human rights defenders.

4. Warning Tool for Human Rights Defenders

While human rights defenders work for the rights of others, they themselves can come under attack from those who want to curtail their activities. Such attacks can be of a violent or a non-violent character. While such attacks – for example the tracking of online activities of human rights defenders or the automated blocking of online content – are addressed regarding AI in a paragraph below, the usage of AI as a warning tool for human rights defenders is also worth discussing.

The matter of using AI as a warning tool for human rights defenders became prominent in the work of the Office of the High Commissioner for Human Rights of the United Nations (OHCHR).³⁹ The duties of the OHCHR encompass the monitoring and the protection of human rights defenders against attacks. The Human Rights Indicators and Data Unit of the OHCHR increasingly focuses on using data to strengthen both monitoring and reporting of threats to human rights and, thereby, also to human rights defenders. It is a difficult undertaking to gather the necessary data for accurate reports on threats to and attacks on human rights defenders.

To find better and more efficient ways to protect human rights defenders the OHCHR entered into a collaboration with

rea of 7 February 2014, UN Doc. A/HRC/25/CRP.1; *Jin Woong Kang*, Human Rights and Refugee Status of the North Korean Diaspora, in: *North Korean Review* 9/2 (2013), pp. 4-17.

³⁸ See *Andreina De Leo*, Fostering Accountability for Human Rights Violations in EU Border Externalization Through the European Ombudsman: The Case of Contesting Financial Support to the Libyan Coast Guard, in: *Journal of Immigrant & Refugee Studies* 23 (2025), pp. 104-120.

³⁹ *Amy Lynn Smith*, Building a Collaboration to Protect Human Rights Defenders, Medium of 16 January 2023, available at: <https://unhumanrights.medium.com/building-a-collaboration-to-protect-human-rights-defenders-26457ae8abd0> (last visited 15 November 2025). The following paragraphs are based on that article.

the US-American AI company Dataminr in April 2022. Dataminr had already worked with the United Nations in the “United Nations Global Pulse” initiative which attempts to bring real-time monitoring and prediction to development and aid programs. The company’s products include “Dataminr Pulse”, a tool to monitor real-time events and to support crisis responses by providing playbooks, messaging tools and post-event documentation.⁴⁰ The “First Alert” platform of Dataminr enables first responders to quickly act in cases of emergencies. It is designed as a real-time critical event discovery solution which is intended to maintain situational awareness and make decisions with confidence.⁴¹

5. Improving the Human Rights Situation on the Ground Directly

AI can also be used by human rights defenders to improve the human rights situation on the ground directly. Individuals or groups who face human rights challenges can be equipped with tools to improve their respective situation based on the capabilities of AI. AI-based speech-to-text applications and translation services are of utmost importance in that regard.

The potential of such AI applications in human rights fieldwork was demonstrated by the Norwegian Refugee Council. The humanitarian organisation used chatbots to help migrants from Venezuela in Colombia. Those chatbots were designed to help

the migrants to learn about their rights under Colombian immigration laws and policies.⁴² AI applications may also be used in humanitarian activities regarding resource management. This is demonstrated by the AI application “Dr. Tania”. This tool was developed by Neurafarm, a farming startup founded by IT experts and engineers from Indonesia, starting in 2018.⁴³ “Dr. Tania” is designed to help Indonesian farmers to tackle crop diseases that threaten their harvest. Thereby, the application also has a link to the right to food. The name of the tool comes from the word “tani” which means “farmer” in Bahasa Indonesia. Farmers can upload a picture of diseased plant to the application. “Dr. Tania” then compares the image of the plant to other pictures in its database and offers the user a diagnosis. In the next step, the farmer is offered information on how to manage the disease and on treatment procedures. The application is designed as a chatbot to ensure easy handling for the user. The uploaded pictures are added to the database, increasing the accuracy of diagnoses in the future. While this is not a usage of an AI application by human rights defenders in a narrow sense it nonetheless shows the potential of such applications

⁴⁰ Dataminr Pulse for Corporate Security, available at: <https://www.dataminr.com/products/pulse/corporate-security/> (last visited 15 November 2025).

⁴¹ Dataminr First Alert, available at: <https://www.dataminr.com/products/first-alert/> (last visited 15 November 2025).

⁴² *Leila Toplic*, AI in the Humanitarian Sector, NetHope of 6 October 2020, available at: <https://nethope.org/articles/ai-in-the-humanitarian-sector/> (last visited 15 November 2025); *Meheret Takele Mandefro*, AI-powered knowledge Chatbot (Norwegian Refugee Council), United Kingdom Humanitarian Innovation Hub of 19 June 2025, available at: <https://www.ukhih.org/news/ai-powered-knowledge-chatbot-norwegian-refugee-council/> (last visited 15 November 2025).

⁴³ *Leander Jones*, Dr Tania: An Indonesian AI Chatbot Helps Farmers Identify and Treat Crop Disease, RESET – Digital for Good of 6 August 2020, available at: <https://en.reset.org/indonesian-ai-powered-app-helps-farmers-identify-crop-disease-05252020/> (last visited 15 November 2025). The following paragraph is based on that article.

in the field of resource management for people affected on the ground.

IV. AI as a Threat to Human Rights Defenders

AI does not only offer great potential for human rights defenders. It can also be a threat to them, as already hinted at above. Human rights defenders can face infringements of their specific rights as human rights defenders for instance by governments using AI. It is known that governments have used AI applications for surveillance of human rights defenders or to block their content in social media platforms.⁴⁴

An example for such usage of AI is the “Oculus” system.⁴⁵ It became known in 2023 that the government of the Russian Federation introduced this system, which helps its authorities to scan the internet for “illegal content”.⁴⁶ The government has stated that the main task of “Oculus” is to recognize violations of Russian laws in pictures and videos on the internet. The system is capable of analysing texts which

are shown within pictures and videos. It was also claimed that the system could evaluate up to 200,000 pictures per day. Before, this task was done manually by employees of the General Radio Frequency Centre, a subordinate regulatory authority of the Federal Service for Supervision of Communications, Information Technology and Mass Media (Roskomnadzor). According to government officials, an employee is able to process 106 pictures and 101 videos per day on average. “Oculus” would also be able to detect “extremist content”, calls for “illegal demonstrations”, “drug-promoting content” as well as “LGBTQ propaganda”. The introduction of “Oculus” and its described usage must be seen as a threat for human rights defenders in Russia. Since the beginning of the full-scale invasion of Ukraine by the Russian Federation in February 2022, the authorities are trying to suppress undesirable opinions even more.⁴⁷ Those activities do not only concern questions of the war, violations of international humanitarian law and human rights within it and its consequences for Russian politics, economy and society, but also the suppression of expressions of LGBTQ rights. Therefore, “Oculus” has the potential to make human rights defenders in Russia turn away from using the internet and especially social media to shed light on human rights issues within the country or in countries, like Ukraine, which are impacted by human rights violations by the Russian state.

Another example concerns Egypt.⁴⁸ The authoritarian government of the North African state uses AI as a tool for political repression and surveillance since the events of the Arab Spring in 2011. AI is used increasingly to act against the free-

⁴⁴ See *Rasma Kaskina/Angelina Cvetkovska*, Artificial intelligence (AI) and human rights: Using AI as a weapon of repression and its impact on human rights, 2024, available at: https://www.europarl.europa.eu/RegData/etudes/IDAN/2024/754450/EXPO_IDA%282024%29754450%28SUM01%29_EN.pdf (last visited 15 November 2025).

⁴⁵ *Ibid.*, p. 3.

⁴⁶ *Matthias Kremp*, Russland automatisiert Suche nach »verbotenen Inhalten« im Internet, Spiegel Netzwelt of 13 February 2023, available at: <https://www.spiegel.de/netzwelt/web/russland-automatisiert-suche-nach-verbotenen-inhalten-im-internet-projekt-oculus-a-ac0d1273-5a71-4035-9a39-757c5d0de9ef> (last visited 15 November 2025). The following explanations are based on that article.

⁴⁷ *Kaskina/Cvetkovska* (fn. 44), p. 3.

⁴⁸ *Ibid.*, p. 4. The following paragraph is based on that report.

dom of expression and the freedom of assembly. Under the Anti-Cyber and Information Technology Crimes Law, which was enacted in 2018, authorities have the legal power to monitor online content and to block websites to protect national security or the economy. While the intended use of the law is to combat political extremists and terrorist groups, in practice it is used against dissidents and other citizens alike. The AI-driven cyber espionage and surveillance techniques of the Egyptian government were and are employed against Egyptian journalists, academics, lawyers, opposition politicians and human rights defenders within the country and those living abroad as well.

V. International AI Legislation and Human Rights Defenders

The specific position of human rights defenders has not been explicitly addressed in recent international legislation concerning the regulation of AI. In the field of human rights, the main instrument at the moment is the aforementioned Framework Convention which was adopted under the auspices of the Council of Europa. The treaty was signed on 5 September 2024. According to its Article 1 para. 1, the object und purpose of the convention is “to ensure that activities within the lifecycle of artificial intelligence systems are fully consistent with human rights, democracy and the rule of law”. As far as can be seen, rights and vulnerabilities of human rights defenders were not discussed during the drafting process. This is not necessarily astonishing as human rights defenders enjoy human rights themselves and therefore can be seen as already being included

in human rights codification at the international level. However, during that process critical remarks concerning the rights of individuals and groups who are engaged in protective activities were made by Michel Frost, the United Nations Special Rapporteur on Environmental Defenders under the Aarhus Convention.⁴⁹ In a statement on the proposed Council of Europe Framework Convention,⁵⁰ addressed to the Committee of Ministers of the Council of Europe and the permanent representatives of the member states of the Council of Europe in Strasbourg, he explicitly criticized the provisions on national security and national defence in the draft convention which are now also present in the adopted instrument.⁵¹ According to Article 3 para. 2 Framework Convention “A Party shall not be required to apply this Convention to activities within the lifecycle of artificial intelligence systems related to the protection of its national security interests, with the understanding that such activities are conducted in a manner consistent with applicable international law, including international human rights law obligations, and with respect for its democratic institutions and processes.” Furthermore, according to Article 3 para. 4 Framework Convention “Matters relating to national defence do not fall within the scope of this Convention.” Frost argues in his statement

⁴⁹ Convention on Access to Information, Public Participation in Decision-making and Access to Justice in Environmental Matters of 25 June 1998, UNTS vol. 2161, p. 447.

⁵⁰ Committee of Minister of the Council of Europe, Draft Framework Convention on artificial intelligence, human rights, democracy and the rule of law, CoE Doc. CM(2024)52-prov1 of 15 March 2024 (Draft Framework Convention).

⁵¹ The statement by *Michel Frost* can be accessed here: https://unece.org/sites/default/files/2024-05/SR_EnvDefenders_Statement_CoE_FrameworkConvention_AI%20and%20Human%20Rights_08.05.2024.pdf (last visited 15 November 2025).

that “(the) complete exemption for matters relating to national defence and the vaguely-worded exemption of matters relating to national security (...) creates a significant risk for abuse and legal loopholes. Indeed, many AI systems used for surveillance and monitoring of the activities of environmental defenders could thereby be excluded from the scope of the Framework Convention on the basis that the activities are allegedly necessary for ‘the protection of national security interests.’”⁵² In line with the recommendations by the Parliamentary Assembly of the Council of Europe,⁵³ he, therefore, called on member states as a matter of absolute urgency “when finalizing the draft Framework Convention, to revise current articles 3 (2) and (4) by limiting the national security interest and national defence exceptions. While such exceptions may be warranted under certain circumstances, a blanket exception for matters of national defence is not. Instead, the text of the Framework Convention must provide in unequivocal terms that AI activities necessary to protect national security interests or national defence must be conducted strictly in line with international human rights law and other international obligations, including article 3 (8) of the Aarhus Convention. This means also that any exception to the rules and principles under the Framework Convention, including in relation to matters of national interest or national defence, must pass the tripartite test of legality, proportionality and necessity under international human rights law.”⁵⁴ While his mandate only concerns environmental de-

fenders under the Aarhus Convention,⁵⁵ the same apprehensions are true for human rights defenders as both are engaged in activities of defending vulnerable conditions and people and often use the same methods and techniques for their work.

VI. Conclusion

The explanations and examples presented in this article have shown on the one hand, that human rights defenders can profit in their work from the usage of AI-driven technologies much in the same ways as individuals, groups and organisations in other work environments. AI technologies can be used to simplify task, finish tasks quicker and add practical and creative potential to human rights work. However, using AI by human rights defenders also has the same limitations and challenges as in other work environments.⁵⁶ There is a necessity to keep humans in the loop for checking flaws in the output of AI technologies. Moreover, for the sake of transparency, AI-created content must be labelled as such to avoid allegations of misconduct or deception.⁵⁷ Furthermore, the necessary data to feed the AI mechanisms

⁵² *Ibid.*, p. 2.

⁵³ Parliamentary Assembly of the Council of Europe, Opinion on Draft Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law, Opinion 303 (2024) of 18 April 2024.

⁵⁴ *Frost* (fn. 51), pp. 2, 3.

⁵⁵ The mandate can be found in Decision VII/9, adopted by the meeting of the parties to the Aarhus Convention, UN Doc. ECE/MP.PP/2021/2/Add.1.

⁵⁶ See also *Sam Bowman*, The role of artificial intelligence in predicting human rights violations, Open Global Rights of 14 November 2024, available at: <https://www.openglobalrights.org/the-role-of-ai-in-predicting-human-rights-violations/> (last visited 15 November 2025).

⁵⁷ See for an example *Luke Taylor*, Amnesty International criticised for using AI-generated images, *The Guardian* of 2 May 2023, available at: <https://www.theguardian.com/world/2023/may/02/amnesty-international-ai-generated-images-criticism> (last visited 15 November 2025).

must be obtained, which can be a challenging process. Depending on legislation at the national and the international level, further legal requirements must be fulfilled. On the other hand, AI technologies are already used by authoritarian governments to hinder human rights work and to persecute human rights defenders. As the capabilities of AI are growing at an ever-faster rate and AI-driven tools and applications become more available, there is no doubt that AI will have a great impact on the work of human rights defenders in the future, to their benefit as well as to their detriment. More scholarly work will be needed to accompany these developments.

Vita

The author is a professor of constitutional law at the University of Applied Sciences for Police and Public Administration in North Rhine-Westphalia, Germany.