Artificial Intelligence (AI) and Human Rights: Using AI as a Weapon of Repression and Its Impact on Human Rights

Technical Report · June 2024

DOI: 10.2861/907329

CITATIONS
23

READS
3,227

1 author:

Hamid Akın Ünver
Özyeğin University
70 PUBLICATIONS **462** CITATIONS

SEE PROFILE

IN-DEPTH ANALYSIS

Requested by the DROI subcommittee

European Parliament

Artificial intelligence (AI) and human rights: Using AI as a weapon of repression and its impact on human rights

Author:
H. Akin ÜNVER

European Parliament coordinator:
Policy Department for External Relations
Directorate General for External Policies of the Union

PE 754.450 – May 2024

EN

DIRECTORATE-GENERAL FOR EXTERNAL POLICIES
POLICY DEPARTMENT

IN-DEPTH ANALYSIS

Artificial intelligence (AI) and human rights: Using AI as a weapon of repression and its impact on human rights

ABSTRACT

This in-depth analysis (IDA) explores the most prominent actors, cases and techniques of algorithmic authoritarianism together with the legal, regulatory and diplomatic framework related to AI-based biases as well as deliberate misuses. With the world leaning heavily towards digital transformation, AI's use in policy, economic and social decision-making has introduced alarming trends in repressive and authoritarian agendas. Such misuse grows ever more relevant to the European Parliament, resonating with its commitment to safeguarding human rights in the context of digital transformation. By shedding light on global patterns and rapidly developing technologies of algorithmic authoritarianism, this IDA aims to produce a wider understanding of the complex policy, regulatory and diplomatic challenges at the intersection of technology, democracy and human rights. Insights into AI's role in bolstering authoritarian tactics offer a foundation for Parliament's advocacy and policy interventions, underscoring the urgency for a robust international framework to regulate the use of AI, whilst ensuring that technological progress does not weaken fundamental freedoms. Detailed case studies and policy recommendations serve as a strategic resource for Parliament's initiatives: they highlight the need for vigilance and proactive measures by combining partnerships (technical assistance), industrial thriving (AI Act), influence (regulatory convergence) and strength (sanctions, export controls) to develop strategic policy approaches for countering algorithmic control encroachments.

EP/EXPO/DROI/FWC/2019-01/Lot6/1/C/30
May 2024 – PE 754.450

EN

Policy Department, Directorate-General for External Policies

AUTHOR

- H. Akin ÜNVER, Associate Professor, Özyeğin University, Turkey

PROJECT COORDINATOR (CONTRACTOR)

- Trans European Policy Studies Association (TEPSA)

This paper was requested by the European Parliament's Subcommittee on Human Rights (DROI).

CONTACTS IN THE EUROPEAN PARLIAMENT

Coordination: Rasma KASKINA, Policy Department for External Relations

Editorial assistants: Balázs REISS and Kristina WILHELMSSON

Feedback is welcome. Please write to poldep-expo@europarl.europa.eu

To obtain copies, please send a request to poldep-expo@europarl.europa.eu

VERSION

English-language manuscript completed in April 2024.

This paper will be published on the European Parliament's online database, 'Think Tank'.

Table of contents

List of abbreviations

ACF	Automated content filtering
ADRN	African Digital Rights Network
AI	Artificial Intelligence
AJL	Algorithmic Justice League
CCP	Chinese Communist Party
CCTV	Circuit Television
CoE	Council of Europe
DPI	Deep packet inspection
DROI	European Parliament's Subcommittee on Human Rights
EEAS	European External Action Service
EP	European Parliament
FBI	Federal Bureau of Investigation
FIDH	International Federation for Human Rights
FSB	Federal Security Service of Russia
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GPAI	Global Partnership on AI
GRU	Main Intelligence Directorate
ICE	Immigration and Customs Enforcement
IDA	In-depth Analysis
ICRG	Islamic Revolutionary Guard Corps
IEE	Institute of Electrical and Electronics Engineers
IP	Internet Protocol
ISP	Internet Service Provider
LLM	Large language models
NIN	National Information Network
NIP	National Internet Project
NLP	Natural Language Processing
OECD	Organisation for Economic Co-operation and Development
SDGs	Sustainable Development Goals
TTPs	Techniques, Tactics and Procedures
UAE	United Arab Emirates
UN	United Nations

UNESCO	United Nations Educational, Scientific and Cultural Organization
USA	United States of America
US	United States
VPN	Virtual Private Network
WEF	World Economic Forum

1 Introduction

Rapid advances in artificial intelligence (AI) are presenting unique challenges to democracy and state-society relations with which the nature of political control is experiencing a significant transformation. AI is a **general-purpose technology** which impacts almost all technological, financial and communication sectors. Hence, the very nature of AI is also altering how power is exercised and maintained by influencing the social and psychological levers that maintain and bolster such power. Such changes directly impact governance across the world, creating newer forms of stress on human rights and fundamental freedoms as more states rely on emerging technologies to amplify control over their societies and data.

This in-depth analysis (IDA) delves into a concerning aspect of modern governance: the use of AI-based tools and techniques for visible, as well as increasingly invisible forms of control and manipulation. Essentially, this involves employing AI systems to monitor, influence and suppress opposition or dissent as well as respective information and data flows, often with high degrees of efficiency and minimal transparency. These methods of control appear in various forms, ranging from expansive surveillance networks or electoral manipulation to more subtle methods of managing information and spreading propaganda online. Such technologies' impact is profound, affecting individual rights and the critical functioning of democratic societies, all of which are expected to become more technically complicated and nuanced over the next decade.

This IDA aims to dissect the increasingly complicated role AI plays in repression and limitations imposed on democratic expression in authoritarian states. It also examines how such technology is being used against the public by an increasing number of countries. The evolution of control has shifted from physical barriers to less visible, yet more invasive digital pathways, using advanced techniques such as deep learning for surveillance, natural language processing for censorship and predictive analytics to anticipate dissident group actions.

In the following sections, this analysis will break down the theoretical, methodological and policy-relevant aspects of repression, identify trends in AI misuse by authoritarian regimes and present case studies demonstrating the real-life effects of these technologies.

Furthermore, support will be provided by this analysis to the European Parliament (EP)'s Subcommittee on Human Rights (DROI) in fulfilling its broad range of responsibilities by:

- **Offering critical insights** into how emerging technologies such as AI can be both a valuable tool for as well as a threat to human rights and democracy. With improved identification of pathways that lead to a healthier use of AI in democracies and the preconditions for misuse through algorithmic authoritarianism, DROI can better assess the effectiveness of existing European Union (EU) instruments in safeguarding against such threats, ensuring they are robust enough to address the challenges posed by AI.
- **Mapping out options** to facilitate more informed and effective dialogues with international human rights organisations. Aligning the EU's approach with global standards and discussions on AI and human rights can help DROI's contribution to a consistent and unified international approach when addressing the challenges posed by AI in the realm of human rights.
- **Serving as a reference document** for DROI's analysis of human rights issues in specific regions or subjects. It can provide crucial data and perspectives for the Subcommittee's reports, particularly as regards countries where algorithmic authoritarianism might be of concern.
- **Assisting in mainstreaming AI-specific human rights language and lexicon** across different organs of the EP, particularly regarding the use and regulation of AI technologies. This

ensures a holistic approach to human rights across all EU policies and actions, guiding policy-makers and officials towards common language, definitions and concepts in analysing the human rights impact of emerging technologies.

Furthermore, the analysis will provide a **crucial overview of AI-driven policy challenges on human rights within the EU,** both through the involvement of Western technology companies within the EU market and the use of EU-originated AI technologies in authoritarian countries, emphasising the point that it is not always the 'usual suspects' that engage in systematic AI repression. This will enable the EP to improve its conception of international partnership and collaboration ecosystems, by employing a more comprehensive set of criteria in detecting and observing algorithmic injustices beyond well-known authoritarian states.

2 Defining repression

Political repression is a classical and well-developed concept in both traditional scientific and policy discourses, with diverse interpretations shaped by different contexts, cases and disciplinary perspectives. While some view repression as overtly violent, others emphasise its subtler manifestations. For policy purposes, operationalised definitions of **'AI-driven repression'** and **'algorithmic repression'** are necessary. Yet, such policy definitions must be derived from more complex terminology work undertaken by the scientific community, which is why the following definitional review aims to clarify the terms' contours by examining various definitions provided by influential works in both fields. It is imperative for policy-makers to understand the nuances of prevalent definitions of algorithmic repression both in academia and policy domains to be able to critique these definitions and alter them as new technologies start influencing state-society relations.

2.1 Scientific definitions

The concept of digital repression, as articulated by Steven Feldstein (2021), encompasses a broad range of state-sponsored activities that exploit **information and communication technologies (ICTs)** as instruments of power to suppress dissent and control populations[1]. This form of repression is an evolution of traditional methods, repurposed for the digital age, where the internet, mobile devices and a vast array of digital tools have become integral to everyday life.

Digital repression involves systematic efforts by state actors to employ advanced technologies in monitoring the digital footprints of individuals and groups. Through surveillance, governments can access a wealth of information that citizens generate online, from their location data to their communication patterns, social networks and even consumer behaviours. This allows states not only to conduct large-scale surveillance but also to coerce populations by threatening exposure and punishment for online activities, as well as using **extrajudicially collected personal data** in legal processes. Beyond surveillance, digital repression includes the manipulation of information to shape public opinion or silence opposition. This can be achieved through the propagation of state-sponsored propaganda, the deployment of bot networks that spread disinformation, and the censorship of online content that is considered subversive or detrimental to the state's narrative.

One of the most insidious aspects of **digital repression** is its capacity to deter activities or beliefs that challenge the state, without necessarily leading to confrontations. The belief that one is constantly being watched can induce self-censorship among citizens, stifling dissent even before it is voiced. This chilling

[1] S. Feldstein, *The rise of digital repression: How technology is reshaping power, politics, and resistance*, Oxford University Press (Oxford: United Kingdom), 2021; S. Feldstein, 'The Road to Digital Unfreedom: How Artificial Intelligence Is Reshaping Repression', *Journal of Democracy*, Vol 30, No 1, 2019, pp. 40-52.

effect on freedom of expression and association is a cornerstone of digital repression, as it subtly transforms the behaviour of individuals through a perceived threat of repercussions for anti-state actions or ideologies. This is a key concern for the EP, as many authoritarian governments dismiss European concerns about digital human rights violations. Whilst there is no overt suppression of information, deeper, more sinister networks of information suppression that generate self-censorship often elide the attention of European monitoring attempts. That said, digital repression can also be more overtly coercive. States may deploy cyber-attacks against opposition groups, manipulate digital platforms to disrupt the organisation of protests or employ legal instruments to justify the arrest and persecution of digital dissidents.

The notion of **'algorithmic repression'** is increasingly critical as it captures the subtle, yet powerful, ways in which technology and social media companies, alongside state actors, can help perpetuate hegemonic control and suppress dissent. Zeynep Tufekci (2017)[2] focuses on the troubling trend that the very algorithms powering social media can clandestinely quash opposition through means such as **algorithmic filtering** and **shadow banning,** imposing a form of digital censorship that is as effective as it is imperceptible. Margaret E. Roberts (2018) dives into the heart of China's sophisticated information control, revealing not just content blocking, but the craft of flooding the digital arena with noise **(content flooding** or **hashtag hijacking)** – a technique that distracts more than it confronts, a subtlety that marks the very essence of digital repression[3].

In her book 'Weapons of Math Destruction', Cathy O'Neil (2017) explores how algorithms, particularly those used in big data, can perpetuate and exacerbate social and economic inequalities, leading to forms of repression – not necessarily confined to autocracies. O'Neil argues that many of these algorithms, while seemingly neutral and objective, are based on biased data or flawed assumptions. This can result in discriminatory outcomes, such as unfairly targeting certain groups for police surveillance, denying individuals' opportunities based on opaque credit scoring systems, or perpetuating hiring biases[4]. O'Neil warns about the **dehumanising effects** of these automated decisions, especially when used by those in power. She highlights how these algorithms, particularly when deployed by political actors, can systematically marginalise certain groups, thereby being manifested as a form of political repression. Safiya Umoja Noble (2018) does not focus on algorithmic repression *per se*, but exposes the insidious biases woven into the fabric of search engines such as Google[5] and reveals a technological infrastructure that not only reflects but also amplifies **racial and gender prejudices.**

Significantly, the earliest and most influential definitions of algorithmic repression focus on platform-level dynamics often seen between users and technology companies or social media platforms in Western democracies. The origin of the terms 'algorithmic repression' or 'algorithmic injustice' in fact originates from a plethora of sources. These concern *inter alia* **the use of automation to cause racial and gender prejudices** in democracies and later expansion of the term to incorporate authoritarian states' use of these tools, as a 'second wave' addition to the original coinage[6].

---

[2] Z. Tufekci, 'Twitter and tear gas: The power and fragility of networked protest', Yale University Press (New Haven: United States of America), 2017. Shadow banning occurs when a social media platform restricts a user's content from showing up without notifying the user, in a hidden and unannounced fashion.

[3] M. Roberts, *Censored: distraction and diversion inside China's Great Firewall*, Princeton University Press (Princeton, New Jersey, United States of America), 2018.

[4] C. O'Neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*, Penguin Random House LLC (New York, New York State: United States of America), 2017.

[5] S. U. Noble, *Algorithms of oppression: How Search Engines Reinforce Racism*. New York University Press (Manhattan, New York State: United States of America), 2018.

[6] P. N. Howard, The digital origins of dictatorship and democracy: Information technology and political Islam, Oxford University Press (Oxford: United Kingdom), 2010.

The academic discourse around algorithmic repression is expanded by more interdisciplinary scholars who examine the multifaceted impact of digital technologies on freedom and control. For instance, Frank Pasquale (2015) provides a critical examination of how algorithmic processes, often proprietary and secretive, govern economic opportunities and the distribution of information in a society[7]. They reveal the **opacity of algorithmic decision-making** in crucial sectors such as finance and media, where the lack of transparency can lead to a form of repression by systematically disadvantaging individuals and groups without any clear recourse. Similarly, Virginia Eubanks (2018) sheds light on the **socio-economic dimensions** of algorithmic decision-making[8] and explores how automated systems are employed in public services, often resulting in a new form of digital divide that exacerbates existing inequalities, marginalising the poor by reinforcing systemic biases in seemingly objective technologies.

Shoshana Zuboff (2023) conceived a highly cited work that expands consideration to the economic underpinnings of digital repression[9]. Zuboff introduces the concept of **surveillance capitalism,** where personal data is commodified and exploited by technology giants, leading to a form of social control that manipulates and modifies behaviour in the service of market objectives, which can be repressive in both intent and outcome. In the same vein, Jack M. Balkin's concept of **'information fiduciaries'** (2014) posits a framework for mitigating the repressive potential of algorithmic systems[10]. By proposing that information custodians – such as social media platforms – should be bound by the same ethical obligations as professionals like doctors and lawyers, Balkin suggests a model where the power of algorithms could be harnessed responsibly, without infringing individual rights.

The term 'AI repression' is thus not widely established with a single, clear definition, but the concept can be derived from the intersection of AI applications and repressive actions.

Key definitional nuances stem from the expressions **'AI versus algorithms'** and **'repression versus authoritarianism'** often being used interchangeably without clear delineations. Broadly speaking, **'AI-driven repression'** is a term that specifically refers to the use of AI technologies to suppress dissent, control populations, or limit freedoms[11]. It implies a direct and active role for AI in repressive actions, often by governments or authoritative bodies. However, **'Algorithmic repression'**[12] can also encompass a broader range of technologies beyond AI, including simpler algorithmic systems that might not qualify as AI. This is also a broader term that can include private sector actions, for instance, social media algorithms' suppressing certain types of content. **'Algorithmic authoritarianism'** is more encompassing than the previous two definitions. It refers to a broader system of governance or control where algorithms play a central role in decision-making processes, surveillance and control mechanisms[13]. It is not limited to outright repression but includes the use of algorithmic tools to maintain and enforce authoritarian governance and control dissidents' information flows, without necessarily engaging in repression. Finally, **'AI-driven authoritarianism'** has a specific focus on AI technologies that exist within the broader

[7] F. Pasquale, The black box society: The secret algorithms that control money and information, Harvard University Press (Cambridge, Massachusetts: United States of America), 2015.
[8] V. Eubanks, Automating inequality: How high-tech tools profile, police, and punish the poor, St. Martin's Press (New York, New York State: United States of America), 2018.
[9] S. Zuboff, *The age of surveillance capitalism*, Routledge (London: United Kingdom), 2023, pp. 203-213.
[10] Jack M. Balkin developed this concept in a series of papers, as early as 2014. J. M. Balkin, 'Information Fiduciaries in the Digital Age', *Balkinization*, 5 March 2014.
[11] J. Earl, T.V., Maher and J. Pan, 'The digital repression of social movements, protest, and activism: A synthetic review', *Science Advances*, Vol 8, No 10, 2022.
[12] N. Ettlinger, 'Algorithmic affordances for productive resistance', *Big Data & Society*, Vol 5, No 1, 2018.
[13] O. Schlumberger, M. Edel, A. Maati and K. Saglam, 'How Authoritarianism Transforms: A Framework for the Study of Digital Dictatorship', *Government and Opposition*, 2023, pp.1-23.

algorithmic ecosystem[14]. It denotes a form of governance or control system where AI is a key tool in maintaining authoritarian rule. This can include surveillance, predictive policing and social scoring systems. However, the use of 'digital' in both 'authoritarianism' and 'repression' contexts broadens the context by referring to **Information and Communications Technologies (ICTs)** and other digital systems such as computers, social media systems, messaging applications and smartphones, sometimes used interchangeably with other definitions above.

The term 'AI repression' is thus not widely established with a single, clear definition, but the concept can be derived from the intersection of AI applications and repressive actions. Below are three interpretations that reflect what could be considered 'AI repression' based on current discussions and concerns in the field:

- **Algorithmic surveillance and censorship:** AI repression can refer to the use of machine learning algorithms and AI by state or corporate actors to conduct surveillance, censor information, and suppress dissent. This might involve using AI to monitor social media, predict protest activities, and flag or remove content that is deemed subversive or contrary to the interests of those in power.
- **Automated decision-making bias:** Another aspect of AI repression can be seen in the systemic biases that are embedded in automated decision-making systems. This refers to algorithms that perpetuate social or political inequalities by marginalising certain groups based on race, gender, or socioeconomic status. The AI systems, through biased data or flawed programming, may reinforce existing power structures and suppress opportunities for the affected groups.
- **Differential technological enforcement:** AI repression might also describe scenarios where AI tools are selectively deployed to target specific populations or individuals, resulting in disproportionate impacts on civil liberties. For instance, AI-powered facial recognition technologies might be used to identify and suppress political activists or minority groups more aggressively than other populations.

It is important to note that these definitions are based on contemporary discussions surrounding the ethics and impact of AI. Accordingly, the term 'AI repression' might evolve or be defined more concretely in future research and policy discussions. However, for this IDA **'algorithmic authoritarianism'** is preferred to capture a broad range of repertoires that can apply to authoritarian states, which go beyond repression by encompassing surveillance and monitoring practices, albeit without necessarily engaging in follow-up action.

2.2 Policy definitions

As real-world implications of algorithmic authoritarianism steer deeper into state-society relations and impact democratic processes and expression of dissent across the world, international bodies are increasingly vocal about the perils of deploying automation in political processes. **Freedom House,** in its Freedom on the Net reports, has been meticulously assessing internet freedom across the globe, raising alarms over practices such as pervasive online surveillance, disinformation campaigns and deliberate internet shutdowns[15]. Particularly insidious are the AI-enabled technologies such as facial recognition, which have been repurposed in some regions to single out and suppress protestors, thereby chilling dissent.

The United Nations (UN) Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression[16] has considered how digital technologies, especially those driven by AI,

[14] K. Crawford, The atlas of AI: Power, politics, and the planetary costs of artificial intelligence, Yale University Press (New Haven, Connecticut, USA), 2021.
[15] A. Funk, A. Shahbaz and K. Veseinsson, 'Freedom on the Net 2023: The Repressive Power of Artificial Intelligence', Freedom House, 2023.
[16] UN Special Rapporteur on freedom of opinion and expression, webpage, nd; United Nations, Our Common Agenda: Report of the Secretary-General, 2021; United Nations, Our Common Agenda: Policy Brief 5. A Global Digital Compact — an Open, Free and Secure Digital Future for All, May 2023.

pose a significant challenge to freedom of speech. They have pinpointed AI-driven surveillance that covertly monitors personal communications, automated content filtering (ACF) that can silently stifle diverse voices and algorithmic biases that can inadvertently reinforce societal prejudices. All such mechanisms can curtail human rights in the digital space. Access Now, a group at the forefront of digital rights advocacy, casts a spotlight on the stark reality of state-sponsored internet shutdowns, flagging them as a blatant form of digital repression. However, beyond these overt acts, they bring to light the subtler, yet equally damaging tactics such as AI-enhanced surveillance that can pinpoint and isolate activists, as well as targeted digital onslaughts aimed at individuals who challenge the status quo[17].

The Council of Europe (CoE), through its Committee on Artificial Intelligence (CAI), is actively working on a Framework Convention on AI, Human Rights, Democracy and the Rule of Law. This convention is being developed to ensure that AI systems are used in a manner that aligns with the Council's corresponding standards. The focus is on ensuring that the application of AI systems does not directly or indirectly undermine democratic processes or endanger human rights. While the Council of Europe does not provide a direct definition of **'AI-driven repression',** its work and goals imply a concern for preventing AI systems from being used in ways that could infringe on human rights or democratic principles. This includes the potential for AI to be used for surveillance, manipulation of information, or other actions that could repress or unduly influence the public in a democratic society[18]. It underscores the importance of transparency, accountability and non-discrimination in AI applications, emphasising that these systems should not be used as instruments for undemocratic control or perpetuation of inequalities[19].

The EP's approach to AI and its potential for misuse in 'algorithmic authoritarianism' is shaped by principles outlined in the proposed **EU AI Act**[20]. It has been underlined that 'No single definition of artificial intelligence is accepted by the scientific community and the term 'AI' is often used as a 'blanket term' for various computer applications'[21]. The Act aims to ensure that AI systems are safe, transparent, traceable, non-discriminatory and environmentally friendly, with human oversight to prevent harmful outcomes. It categorises AI systems that present unacceptable risks and will be banned, such as those capable of cognitive behavioural manipulation, social scoring and real-time remote biometric identification systems such as facial recognition. Furthermore, AI systems that could negatively affect safety or fundamental rights are classified as high-risk and subject to stringent regulations[22].

The Act considered past definitions of AI that were proposed by the High-Level Expert Group on AI in March 2021 and the European Commission's Joint Research Centre's attempt to come up with an operational AI definition, categorising various AI subdomains from political, research and industrial viewpoints. However, considering both definitions to be inadequate and recognising the need for a more precise description of an AI system for legal clarity in the new AI framework, the **European Commission** suggested introducing a clear legal definition in EU law. This proposal was largely aligned with one used by the **Organisation for Economic Co-operation and Development (OECD),** which defines an AI system as software created using

---

[17] See for example, 'Ban Biometric Surveillance. Access Now', webpage, nd.

[18] Council of Europe, 'CAI – Committee on Artificial Intelligence', webpage, nd.

[19] Council of Europe, Algorithms and Human Rights: Study on the human rights dimensions of automated data processing techniques and possible regulatory implications, 2018; Council of Europe, 'Council of Europe and Artificial Intelligence', 2023; Council of Europe Revised Zero Draft [Framework] Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law, Committee of Artificial Intelligence, (2023)01, 6 January 2023.

[20] European Parliament press release "Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI", 9 December 2023; European Council Press Release, 'Artificial intelligence act: Council and Parliament strike a deal on the first rules for AI in the world'. 9 December 2023;.

[21] T. Madiega, 'European Parliament Briefing on EU Legislation on Artificial Intelligence', PE 698.792, June 2023, pp. 3-4.

[22] European Parliament, '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', P9_TA(2023)0236. 14 June 2023.

specific techniques and approaches, capable of generating outcomes such as content, predictions, or decisions that impact their environment, based on human-defined objectives. However, this is also a working definition, given that there are non-AI systems capable of generating outcomes based on human-defined objectives. These include more classical approaches using statistics, such as regression analysis.

Annex 1 of the proposal lists current techniques and approaches used in AI development, defining an AI system as a spectrum of software technologies, including machine learning, logic and knowledge-based systems as well as statistical methods. This comprehensive definition covers AI systems used either independently or as part of a product. The legislation is designed to be adaptable, covering both present and future AI technologies. Furthermore, Article 3 provides extensive definitions for terms such as 'provider', 'user', 'importer' and 'distributor' of AI systems, applicable to both public and private entities, including definitions for 'emotion recognition' and 'biometric categorisation'.

These new amendments signal a careful approach to the issue of banning and restricting critical data points that feed the most prevalent cases of AI misuse, *inter alia*: biometric categorisation; facial image scraping; CCTV use; emotion recognition; and social scoring. The restrictions consider both the 'first-wave' **(uncontrolled algorithmic bias)** and 'second-wave' **(deliberate algorithmic targeting)** definitions of AI repression.

Despite the prevalent policy definitions discussed above, though, any accurate capturing of the term's breadth remains elusive. Defining 'algorithmic authoritarianism' is indeed elusive and challenging, especially when applied to different contexts. This stems from the inherent complexity and evolving nature of algorithms themselves, as well as the diverse contexts in which they are being applied. Indeed, 'repression' can itself be interpreted differently in various cultural, social and legal contexts. What one group considers repressive, another might view as regulation or necessary oversight. This subjectivity adds another layer of complexity in defining and understanding algorithmic authoritarianism. Furthermore, it is important once again to emphasise that the 'first wave' scientific literature which coined the term has referred to a very different repertoire of cases and contexts compared to the current mainstream EU definition.

Most EU reports and statements on the topic primarily focus on autocratic regime types. However, the original terminology appeared in a different context, referring to the platforms' hidden architectures and automation systems that create gender, racial and ethnic inequalities primarily in Western democracies. Hence, EU-level definitions of algorithmic authoritarianism could avoid any political bias based on countries' regime types and instead focus on specific **techniques, tactics and procedures (TTPs)** when referring to AI-based repressive practices. It is important to note that algorithms become 'authoritarian' when they are used for authoritarian and repressive purposes, rather than when the employing government is listed as 'authoritarian' in regime-type indexes.

3 Case studies on algorithmic authoritarianism and surveillance

3.1 Case and time selection criteria

Algorithmic authoritarianism practices are becoming more common, not just among authoritarian countries, but also democracies, an increasingly worrying global trend. To that end, a full and exhaustive case analysis of all examples is not possible within the confines of this report. The IDA will instead focus on those countries that appear increasingly more frequently within EU statements and reports, which hence are 'cases of concern'. These countries impose significant government control over digital channels, from internet service providers (ISPs) to online platforms, underpinned by extensive surveillance systems. Such

control is pivotal in understanding the mechanisms and scope of algorithmic/AI-based authoritarianism/ repression[23].

Another critical factor is the prevalence of unregulated (or under-regulated) advanced technologies used by authoritarian states in employing repression or information suppression in foreign countries, especially in democracies, including, but not limited to certain EU countries. Over the last decade, the use of emerging technologies by authoritarian states in **Foreign Information Manipulation and Interference**[24] and repression are well-documented, receiving most of the EU's attention in official documents and reports. Additionally, the legal and regulatory frameworks within these nations not only enable, but also often legitimise the use of algorithms for repression, control and censorship. This includes legislation about cybersecurity, information control and national security, offering a legal backdrop to these practices. In most of these authoritarian states, AI and algorithms are used in a deliberately under-regulated fashion, justified within the pretext of combating terrorism or curbing radicalisation[25].

While acknowledging that algorithmic authoritarianism is a widespread concern, the focus of this IDA on Russia, China and Iran is due to their particularly notable roles in this domain and recurring emphasis in EU official documents. However, to present a more nuanced picture and geographical variance, additional cases of Middle-Eastern, North African and Sub-Saharan contexts have also been added here to provide a greater perspective.

Although not immediately relevant to the EU, AI-based manipulation and repression efforts are also becoming increasingly more visible in Latin America. The Venezuelan government under Nicolás Maduro, for example, is reportedly using AI-generated newscasters to spread disinformation[26]. A study by the NGO Cazadores de Fake News found that the virtual journalists, Daren and Noah, were created using Synthesia software, delivering English newscasts that exclusively favour the regime. These avatars come from a catalogue of over 100 multiracial faces provided by Synthesia, which allows users to generate scripted content in multiple languages[27]. Synthesia's service, costing about USD 30 a month, requires no video creation expertise and can synchronise scripts with avatars in over 100 languages. Similar AI-based information manipulation efforts have been identified around the election periods of, *inter alia*, Bangladesh, Türkiye, Brazil and Argentina over the last year[28].

To provide a balanced view of the technical, policy and case-specific nuances of algorithmic authoritarianism, some topics or technologies are consciously omitted from this IDA to focus on significant and influential global actors in the field of AI (other techniques can be found in the Annexes). The interplay between politics and AI, particularly in the context of algorithmic authoritarianism, remains a complex and under-researched area with significant variations across different geopolitical landscapes. In regions such as Sub-Saharan Africa or certain parts of Southeast Asia and Latin America, there is a dearth of comprehensive studies on how AI is leveraged for political means.

---

[23] M. Crosston, 'Cyber colonization: The Dangerous Fusion of Artificial Intelligence and Authoritarian Regimes', *Cyber, Intelligence, and Security Journal*, Vol 4, No 1, 2020, pp. 149-171.

[24] European External Action Service, '1st EEAS Report on Foreign Information Manipulation and Interference Threats', 7 February 2023.

[25] R. S. Andersen, 'Video, algorithms and security: How digital video platforms produce post-sovereign security articulations', *Security Dialogue*, Vol 48, No 4, 2017, pp. 354-372.

[26] M. L. Paul, 'Noah' and 'Daren' report good news about Venezuela. They're deepfakes', *The Washington Post*, 2 March 2023.

[27] J. Daniels and M. Murgia, 'Deepfake 'news' videos ramp up misinformation in Venezuela', *Financial Times*, 17 March 2023.

[28] B. Parkin, 'Deepfakes for $24 a month: how AI is disrupting Bangladesh's election', *Financial Times*, 14 December 2023; D. Ioannou, 'Deepfakes, Cheapfakes, and Twitter Censorship Mar Turkey's Elections', *Wired*, 26 May 2023; M. Margolis, R. Muggah, 'Brazil's fake-news problem won't be solved before Sunday's vote', 27 October 2022; D. Feliba, 'How AI shaped Milei's path to Argentina presidency', *The Japan Times*, 22 November 2023.

The decision to focus this IDA on the past decade is anchored in the pivotal developments that have shaped the landscape of algorithmic authoritarianism in Russia, China and Iran during this time. The period has been characterised by significant technological advancements, particularly in AI, machine learning and big data analytics, which are instrumental in the evolution of state surveillance and control mechanisms. The last ten years have also witnessed a profound increase in digital surveillance, marking a shift from traditional means to sophisticated, algorithm-driven approaches[29].

This era is crucial not only for understanding how these technological advancements have been harnessed by governments to enhance their repressive capabilities but also for considering the simultaneous global proliferation of the internet and exponential rise in social media usage, transforming these platforms into new battlegrounds for information control and manipulation. The role of digital channels in both expression and repression has become increasingly prominent, offering new means for governments to monitor and control the flow of information. This period has also seen notable legal and policy developments in these countries, with the introduction of new laws and regulations that significantly impact digital rights and freedoms. These changes provide a crucial backdrop for understanding the mechanisms and extent of algorithmic authoritarianism.

3.2 Advanced algorithmic control at scale: The case of China

3.2.1 Xinjiang region and the persecution of Uighur minority

One of the most challenging and well-documented instances of China using AI-based technologies for algorithmic authoritarianism is in Xinjiang, with its targeting of the Uighur Muslim minority. Through comprehensive surveillance apparatus, the Chinese authorities have employed a wide range of technologies, from facial recognition to predictive policing, which is being used to monitor, control and detain a significant portion of this population.

The Xinjiang Uighur Autonomous Region holds considerable importance for China, both historically and in terms of its resources. Predominantly inhabited by the Uighur Muslim minority, the region has seen a growing influx of Han Chinese residents, a dynamic that has fuelled ongoing tensions. At the heart of the conflict are cultural clashes, restrictions on religious practices and economic inequalities, alongside a strong desire among the Uighurs to preserve and recognise their distinct identity. Since the inception of the People's Republic of China in 1949, the government has pursued policies aimed at cultural assimilation and economic progress. However, these measures are often perceived by the Uighurs as a threat to their cultural heritage. Complicating matters further is the influence of global jihadist movements, which have occasionally garnered support in the region, increasing Beijing's concerns and consequently leading to more stringent security actions in Xinjiang.

Beijing's strategy in the region has evolved from containment to active information suppression, aiming to micro-manage the biometric and personal data of Xinjiang citizens. As part of this shift, a deliberate decision has been taken to harness some of the most advanced AI applications to test and scale some of the surveillance and monitoring advantages of automated systems. The conception of this policy can be traced to the synthesis of two primary streams of thought within the Chinese leadership. Firstly, the global rise of AI presented an irresistible tool that could offer unparalleled surveillance capabilities. Secondly, the unique challenges in Xinjiang required a solution that was pervasive, discreet and pre-emptive. In this milieu, the decision to deploy AI was not an impulse, but a culmination of meticulous planning, reinforced

[29] J. Zeng, 'Artificial intelligence and China's authoritarian governance', *International Affairs*, Vol 96, No 6, 2020, pp. 1441-1459.

by the belief in technological supremacy as a means of governance[30]. The envisioned AI apparatus would not just augment existing surveillance, but also act as a cornerstone for a sophisticated predictive policing programme. This foresaw the utilisation of vast data point collection infrastructure investments, from biometrics to behavioural patterns, fuelling algorithms designed to flag potential dissidents before any overt acts of defiance could be manifested[31]. Consequently, the decision to deploy these advanced tools in Xinjiang was identified as a strategic imperative under the guise of counterterrorism and social stability. This marked a pivotal moment in the trajectory of Chinese surveillance policies, signifying a turn towards an era where algorithmic governance began to overshadow traditional, human resources-oriented mechanisms.

As a result, many data collection initiatives have been launched[32]. One prominent example is the **Sharp Eyes** programme, launched in 2015, which expanded on the **Skynet** initiative started in 2005 for urban surveillance. Sharp Eyes leverages a wide range of data sources, including surveillance cameras, vehicle and license plate recognition cameras, together with virtual identities such as MAC addresses and phone numbers. This data is then integrated using geographic information systems and sent to 'societal resource integration platforms', which are present in various provinces, including Xinjiang. China's data-fusion programmes target specific social groups, especially those considered to be 'focus personnel', such as individuals petitioning the government, those involved in terrorism, or others deemed to be undermining social stability. The Uyghur ethnic minority in Xinjiang is subjected to intense surveillance under these programmes. Tools such as the Integrated **Joint Operations Platform (IJOP)** in Xinjiang link individuals' government-issued ID cards to physical characteristics and monitor for behaviours considered indicative of potential social instability. Moreover, Chinese laws mandate cooperation between private firms and state security organs. These include the 2016 cybersecurity law, the 2017 national intelligence law and the 2021 data security law. Such an environment of increasing rigidity and centralisation places significant emphasis on political stability and necessitates data sharing with government authorities.

Residents' daily movements have become part of a **systematic surveillance ritual,** marked by frequent verification at **checkpoints** or data collection stations. These stations were set up to gather personal data, a process that typically involved scanning identification documents, facial recognition and examining personal communication devices. These checkpoints serve two functions, underscoring the constant presence of state surveillance and collecting detailed data necessary for advanced predictive policing systems. Predictive policing algorithms were developed directly from this extensive data collection and have facilitated the detection of patterns indicating potential dissent or non-conformity. Their objective has not been to address existing crimes, but to predict potential security threats, thus marking a shift from conventional policing methods towards a governance model focused on managing and mitigating risk pre-emptively.

As a result of this data-centric security strategy, many Uighurs have been caught in the predictive policing system's net and subsequently confined to re-education camps. These detentions have been characterised by a lack of transparency, often conducted without formal charges or legal process, based on the ambiguous outcomes of AI system analyses. The **extensive surveillance network** has fostered a sense of

[30] L. Oztig, 'Big data-mediated repression: a novel form of preemptive repression in China's Xinjiang region', *Contemporary Politics*, 2023, pp. 1-22.

[31] J. Leibold, 'Surveillance in China's Xinjiang region: Ethnic sorting, coercion, and inducement', *Journal of contemporary China*, Vol 29, No 121, 2020, pp. 46-60.

[32] S. Kam and M. Clarke, 'Securitization, surveillance and 'de-extremization'in Xinjiang', *International Affairs*, Vol 97, No 3, 2021, pp. 625-642; G, Roche and J. Leibold, 'State Racism and Surveillance in Xinjiang (People's Republic of China)', *The Political Quarterly*, Vol 93, No 3, 2022, pp. 442-450.

constant observation within the Uighur community, resulting in widespread self-censorship and behavioural changes. Routine practices have been discontinued and conversations have become cautious, with people adjusting to the unspoken boundaries set by the surveillance system. This has had a **significant psychological impact on citizens,** as their mere awareness of being observed has fundamentally altered community dynamics.

As a by-product of this digital authoritarianism, a burgeoning economic market for surveillance technologies has emerged. Chinese firms specialising in advanced surveillance hardware and software, such as **Hikvision** and **Dahua Technology,** have found themselves at the centre of a growing domestic market[33]. These companies have reaped substantial profits from government contracts and their technologies have become a testament to the state's ability to control and manage its citizens. The scale of deployments in Xinjiang propelled these firms to the forefront of the global surveillance technology sector, even as they faced international scrutiny and sanctions.

The Chinese government's internal narrative has portrayed the surveillance and re-education measures in **Xinjiang** as necessary for **combating extremism and fostering economic development.** This perspective has been widely disseminated and generally accepted among the Han Chinese majority, due in part to the government's significant control over the domestic information environment. State media has highlighted the development of infrastructure and job creation in the region, thus helping to bolster support for government policies among the general population. Discussions about the impact of these measures on the Uighur population and other ethnic minorities have largely been suppressed, resulting in a skewed public understanding within China[34].

In stark contrast to domestic approval, the **international response** has been **marked by severe criticism.** Investigative reports by journalists, alongside campaigns by human rights organisations, have shone a light on the conditions within the re-education camps and the extensive surveillance apparatus, prompting a wave of international condemnation. Several Western governments, international bodies and advocacy groups labelled the actions of the Chinese government as severe human rights abuses[35]. This led to sanctions being imposed against Chinese officials and technology companies involved with surveillance and repression in Xinjiang, indicating a robust, albeit complicated, international response.

Confronted with international criticism, the Chinese government has responded by consistently **refuting allegations of human rights violations.** Officials often have described the camps as vocational training centres, arguing that the surveillance measures are essential for maintaining stability and combating terrorism. This stance has been accompanied by a strong diplomatic effort to counteract negative portrayals, including inviting select foreign visitors to tour facilities in carefully staged visits aimed at showcasing the government's narrative.

Despite the international scrutiny and pressure, the sophisticated surveillance network in Xinjiang persists. While there have been reports suggesting a decrease in the re-education camps' population[36], the long-term impact and the fate of those who have been detained remain the subject of international concern. Furthermore, the technology developed and perfected in Xinjiang is not only still in use, but is

[33] J. Pan, 'How Market Dynamics of Domestic and Foreign Social Media Firms Shape Strategies of Internet Censorship', *Problems of Post-Communism*, Vol 64, No 3, 2017, pp. 625-642.
[34] S. Kam and M. Clarke, 'Securitization, surveillance and 'De-extremization' in Xinjiang', *International Affairs*, Vol 97, No 3, 2021, pp. 625-642.
[35] S. R. Roberts, 'The biopolitics of China's "war on terror" and the exclusion of the Uyghurs', *Critical Asian Studies*, Vol 50, No 2, 2018, pp. 232-258.
[36] L. Gambino, 'Like a war zone': Congress hears of China's abuses in Xinjiang 're-education camps', *The Guardian*, 24 March 2023.

also being marketed to other countries[37]. This raises alarms about the export of such surveillance capabilities and the potential for other governments to adopt similar methods of social control.

3.2.2 China's social credit system: A technical examination

The **Social Credit System (SCS)** in China, often mischaracterised as a single, monolithic score for each citizen, is an iterated web of interconnected initiatives, policies and technologies aimed at shaping individual and corporate behaviour. At its core, the SCS amalgamates social and behavioural digital data and AI to rank citizens as well as corporations, rewarding or punishing them based on a variety of metrics.

The SCS is **not just a single nationwide system,** but comprises many pilot projects run by city municipalities and even private companies. Over time, the intention is to integrate these projects into a more cohesive national framework, each being designed to address specific areas of social and economic behaviour[38]. For example, one project might focus on financial credibility, tracking loan repayments and financial fraud, while another might emphasise social decorum, monitoring behaviour patterns such as adherence to traffic rules or public interactions[39]. The scale and population size in China, coupled with regional cultural differences present challenges which necessitate tailored approaches to social credit. What works in an urban environment such as Shanghai might not be suitable for a rural area in Yunnan. Each regional system, therefore, includes **scoring variables based on local customs**, **needs and governance styles,** creating varied weightings intended to address regional and cultural contexts[40]. The eventual goal for the Chinese Communist Party (CCP) is to integrate these diverse systems into a national framework, yet the integration process is complex, as it involves reconciling different scoring methods, data standards and policy priorities. The challenge lies in creating a unified system that respects local differences while adhering to national standards[41].

In the SCS ecosystem, **government and private companies** operate in a **complementary fashion.** The government relies on private sector innovation and agility in handling massive datasets, while companies benefit from the legitimacy and regulatory framework the government provides. This symbiosis aims to foster a more efficient and responsive credit system. Companies such as Alibaba and Tencent have vast quantities of data on consumer behaviour[42]. By developing their credit scoring systems, these companies not only serve their commercial interests but also contribute to the broader social credit initiative. Their systems can serve as testing grounds for new approaches that may inform the national SCS's evolution. Aligning the different private systems with the government's SCS presents challenges in terms of data privacy, user consent and standardisation of metrics[43]. There is also a need to manage potential conflicts between companies' profit motives and the government's social objectives.

The **scope of data collection extends beyond basic public records** to include financial transactions, health records, employment status and compliance with civil duties. This information provides a multidimensional profile of a citizen's public, private and financial life. Citizens can also contribute data points, either voluntarily or as part of required interactions with government services. For instance, submitting

---

[37] S. C. Greitens, 'Dealing with demand for China's global surveillance exports', *Brookings Institute*, April 2020.
[38] R. Creemers, 'China's Social Credit System: an evolving practice of control', 2018.
[39] F. Liang, V. Das, N. Kostyuk, and M. M. Hussain, 'Constructing a data-driven society: China's social credit system as a state surveillance infrastructure', *Policy & Internet*, Vol 10, No 4, 2018, pp. 415-453; D. Mac Síthigh and M. Siems, 'The Chinese social credit system: A model for other countries?', *The Modern Law Review,* Vol 82, No 6, 2019, pp. 1034-1071.
[40] N. Kobie, 'The complicated truth about China's social credit system', *Wired*. 7 June 2019.
[41] Z. Yang, 'China just announced a new social credit law. Here's what it means', MIT Technology Review, 22 November 2022.
[42] Congressional Research Service, 'China's Corporate Social Credit System', IF11342, 17 January 2020.
[43] D. M. Síthigh and M. Siems, 'The Chinese social credit system: A model for other countries?', *The Modern Law Review*, Vol 82, No 6, 2019, pp. 1034-1071.

information for licenses, social services, or even participating in community activities can be tracked and factored into the system. There is an intersection between various sectors in data sharing, such as the cooperation between banking, transportation and law enforcement. This creates an extensive network for data mining, where patterns of behaviour can be observed across different aspects of daily life[44].

Surveillance extends **into public transportation,** with systems designed to monitor individuals' compliance with regulations in real time. For example, payment evasion on public transit may be detected and logged as a negative action. Some regions have implemented systems where peers and community members can report both negative and positive behaviours. These peer-reported activities may also be factored into the SCS, thus integrating a dystopian societal watch mechanism that complements technological surveillance. Although the government emphasises positive uses of this system, such as promoting environmental responsibility by rewarding certain behaviours, for instance, recycling and energy conservation, the overall impact of this system on society is mixed. While some Chinese citizens report greater satisfaction with governance and security, others report concern due to the invasive and crowdsourced nature of SCS[45].

The key to the SCS's influence lies in its ability to correlate data from various sources. For instance, a person's online behaviour can be linked with recorded physical activities to present a holistic view. The integration process involves **sophisticated algorithms and continually refined machine learning models** that analyse vast datasets to identify patterns, make predictions and generate scores. Achieving interoperability between different systems and datasets is a technical challenge, involving data format standardisation, synchronisation of update cycles and resolution of discrepancies between various sources. Integration also includes the **creation of feedback mechanisms,** where the outcomes of certain behaviours influence not just the immediate scoring, but also the system's parameters in a dynamic evolution. As data is integrated, questions arise about the **legal frameworks** that govern data protection, sharing, and individual rights. There is an ongoing debate about how to ensure ethical practices are upheld within this extensive data collection and integration process. These scores are not arbitrary figures but are the outcome of a meticulous process where every action is weighted according to its societal value[46].

A financial fraud, for instance, casts a longer shadow on one's social credit than a minor traffic infraction, mirroring the values that the system is designed to uphold. Yet, it is not merely a backwards-looking tally; the system delves into **dimensional analysis, painting a multifaceted portrait of behaviour** that encapsulates civic responsibility, financial probity and social trustworthiness. There is a palpable tension between the system's opacity and growing calls for transparency and fairness, particularly around how one can rectify or challenge the omnipotent algorithm's verdict.

3.2.3 Key actors and entities involved in Chinese algorithmic authoritarianism

Actors directly involved and bearing the most consequences from China's AI authoritarianism are the CCP, central government agencies and provincial administrations.

The **CCP**, China's central authority in the political hierarchy, actively directs and shapes the nation's policy and governance systems, including the strategic deployment of AI[47]. The CCP's governance philosophy, which prioritises social stability and control, has found a powerful instrument in AI technology, which it leverages for a variety of functions aimed at preserving the status quo and suppressing dissent. AI is woven

[44] C. Liu, 'Multiple social credit systems in China', *Economic Sociology: The European Electronic Newsletter*, Vol 21, No 1, 2019, pp. 22-32.
[45] L. C. Backer, 'China's Social Credit System', *Current History*, Vol 118, No 809, 2019, pp. 209-214.
[46] K. L. X. Wong and A. S. Dobson, 'We're just data: Exploring China's social credit system in relation to digital platform ratings cultures in Westernised democracies', *Global Media and China*, Vol 4, No 2, 2019, pp. 220-232.
[47] J. Leibold, 'Surveillance in China's Xinjiang region: Ethnic sorting, coercion, and inducement', *Journal of contemporary China*, Vol 29, No 121, 2020, pp. 46-60.

into the fabric of China's national strategy, as evident from the remarks and directives of President Xi Jinping, who has consistently highlighted the importance of high technology in governance. Under his leadership, AI has been posited not just as a catalyst for economic prowess, but also as a crucial enabler of political governance. The CCP views AI as a means of achieving its political ends, a perspective that is reflected in policy formulations and the party's approach to internal as well as international issues.

Several **central government agencies** play key roles in the development and deployment of AI. The Ministry of Public Security is a critical component of China's state apparatus, tasked with law enforcement, social control, and surveillance. It has been a pioneer in embracing AI for security purposes[48]. Across China, and particularly in regions such as Xinjiang where the government faces ethnic tensions and seeks to pre-empt dissent, the Ministry of Public Security has deployed a vast network of surveillance tools, which serve as a hub of biometric data collection and provision mechanism for other agencies. The **Integrated Joint Operations Platform** is an AI system used in Xinjiang that aggregates data from multiple sources, including closed circuit television (CCTV) cameras and checkpoints, to identify and flag 'suspicious' behaviour that may prompt a police response. It is tasked with collecting and disseminating raw data, as well as analytics associated with such data forms.

The **State Internet Information Office** also known as the Cyberspace Administration of China is the primary gatekeeper of China's cyberspace. Its mandate is broad, including online content regulation, internet security, and digital policy enforcement[49]. The Office holds a crucial role in operationalising China's approach to cyber sovereignty, including c**ontent monitoring** to **monitor** and analyse the vast swathes of data flowing through the Chinese internet, flagging and removing **content** deemed inappropriate or threatening to state interests, as well as c**ensorship and propaganda** to influence public opinion through the promotion of certain content, while suppressing dissenting views, effectively guiding the narrative in a direction favoured by the CCP.

The **National Development and Reform Commission,** China's macroeconomic management agency, plays a strategic role in the country's AI ambitions. It oversees and coordinates economic and social development policies, for instance by integrating AI into the national economy[50]. It formulates plans for the advancement of AI, focusing on areas such as machine learning, data analytics and the broader integration of AI technologies into various sectors of the economy. It also implicitly guides how AI is to be utilised for social governance, ensuring that technological development serves the political and social objectives set by the CCP.

China's utilisation of AI reflects a dual approach where technology serves to propel economic modernity alongside fortifying state control mechanisms. As the central government carves out the broad strokes of the country's AI strategy, local governments bring these visions into reality, often through partnerships with private technology giants. At the forefront, cities such as Hangzhou represent the vanguard of China's smart city aspirations. Its **'City Brain'** project, developed with **Alibaba,** is a testament to this ambition. While the stated purpose is to streamline city management, such as optimising traffic flow, it doubles as a sophisticated surveillance system[51]. It leverages AI to process data from a network of cameras and sensors, effectively tracking human movement and behaviour. This has raised concerns about the extent to which

---

[48] Z. Su, A. Cheshmehzangi, D. McDonnell, B. L. Bentley, C. P. Da Veiga, and Y.T. Xiang, 'Facial recognition law in China', *Journal of Medical Ethics*, Vol 48, No 12, 2022, pp. 1058-105; G. King, J. Pan, and M. E Roberts, 'How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument', *American Political Science Review*, Vol 111, No, 2017, pp. 484-501.

[49] R. Hou, 'Neoliberal governance or digitalized autocracy? The rising market for online opinion surveillance in China', *Surveillance & Society,* Vol 15, No 3 and 4, 2017, pp. 418-424.

[50] H. Roberts, J. Cowls, J. Morley, M. Taddeo, V. Wang and L. Floridi, 'The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation', *AI & society*, Vol 36, 2020, pp. 59-77.

[51] M. Toh and L. Erasmus, 'Alibaba's 'City Brain' is slashing congestion in its hometown', *CNN Business*, 15 January 2019.

monitoring is used beyond public service optimisation, potentially for the suppression of dissent and surveillance of citizens beyond legal and ethical bounds[52].

In Xinjiang, the regional government's implementation of AI surveillance targets the **Uighur population** under the guise of counterterrorism and anti-extremism measures. Here, AI extends its reach into **biometric data, including facial recognition technology and DNA sampling,** intertwined with data from mandatory applications and police checkpoints[53]. This has created a near-inescapable net of surveillance that severely impinges on the Uighurs' freedom and privacy, drawing international condemnation over human rights violations.

Shenzhen's foray into a SCS is another embodiment of AI's penetrating influence on social governance. The **system scores citizens on a spectrum of behaviour,** with high scores bringing benefits such as easier access to credit and lower scores leading to restrictions such as slower internet speeds or bans on travel[54]. The use of AI to process these scores automatically, considering vast amounts of personal data, raises profound questions about consent, data security, and the potential for governmental abuse.

Chongqing, another example, has an extensive network of cameras, many equipped with AI capabilities. This creates an omnipresent monitoring apparatus that not only deters crime but also fosters an environment where **citizens may be reluctant to express dissent or engage in activities** outside the state's normative expectations, due to fears of being watched and judged by an invisible, algorithmic authority[55].

These regional implementations are not isolated, but form a collective mosaic of China's approach to AI as **a tool for comprehensive social management.** The technical prowess of AI systems is impressive, but also unsettling given its breadth and depth. China's application of AI has indeed led to significant improvements in efficiency and public administration, but the cost has been the establishment of an environment where privacy is diminished and the state's ability to control and influence citizens' behaviour is unprecedented.

3.2.4 Transnational involvement of China's AI policy and misuse

China's engagement in transnational AI repression is a manifestation of its broader strategy to control information, suppress dissent and influence debate both domestically and internationally. By harnessing emerging technologies, **China can project its censorship and surveillance capabilities across borders,** affecting not only Chinese nationals living abroad but also foreign citizens and entities that may engage with issues sensitive to the Chinese government.

The most prominent example of China's transnational control is the 'Great Firewall', which has been adapted to affect content globally, making it not just a defensive precautionary system, but a tool of global surveillance. While not an AI system itself, the Great Firewall contains AI-based sub-components such as ACF, algorithmic cybersecurity protocols and content monitoring practices. Beijing pressures international companies and online platforms to **censor and alter content** – both manually and algorithmically – that contradicts its narrative or political agenda, often using its market power as leverage. This often leads to the suppression of topics deemed sensitive by the Chinese government at scale, using automated text detection and **large language model (LLM)** protocols. Apple, for instance, has faced criticism for removing

[52] K. Grön, Z. Chen, and M. Ruckenstein, 'Concerns with Infrastructuring: Invisible and Invasive Forces of Digital Platforms in Hangzhou, China', *International Journal of Communication*, Vol 17, 2023, pp.17-23.
[53] See Lee-Wee S. and P. Mozur, 'China Uses DNA to Map Faces, With Help From the West', *The New York Times*, 3 December 2019.
[54] F. Liang, V. Das, N. Kostyuk and M. M. Hussain, 'Constructing a data-driven society:China's social credit system as a state surveillance infrastructure', *Policy & Internet*, Vol 10, No 4, 2018, pp. 415-453.
[55] M. Keegan, 'Big Brother is watching: Chinese city with 2.6m cameras is world's most heavily surveilled', *The Guardian*, 2 December 2019.

applications from its Chinese App Store that enable users to bypass internet censorship, ostensibly in compliance with Chinese laws[56]. LinkedIn also previously complied with Chinese regulations by censoring certain profiles and content from being viewed in China[57]. This was undertaken both in a manual, case-by-case fashion, as well as algorithmically, as the Great Firewall contains machine learning classifiers which are capable of learning censorship and content removal protocols.

China **exports sophisticated surveillance technologies,** including facial recognition systems and internet monitoring tools. These systems are often embedded with AI capabilities, enabling real-time analysis and identification of individuals, as well as the monitoring of digital communications on a massive scale. For instance, China's technology giant Huawei has helped Ecuador deploy its **ECU-911 system,** an integrated surveillance system spread across the country, raising concerns about the potential for Chinese-style surveillance tactics to spread abroad[58]. Chinese companies have sold surveillance systems, including facial recognition technology to countries such as Zimbabwe and have been involved in the development of smart cities, which include extensive surveillance infrastructure[59].

Chinese state-backed **hackers** have been involved in numerous cyber operations targeting dissidents, ethnic minorities and foreign governments. These operations frequently involve the deployment of AI to automate data theft, carry out sophisticated phishing attacks and analyse large sets of compromised data for valuable intelligence[60]. Uncovered by security researchers, this global cyber espionage campaign targeted managed service providers to steal data from their clients. It was attributed to Chinese state-sponsored actors. There have also been many indictments and accusations from the United States of America (USA) and other so-called Western countries that hackers from China have stolen intellectual property from companies, a task made more efficient by AI algorithms that can sift through large volumes of data rapidly.[61]

AI tools are employed to tailor and **disseminate propaganda** across various digital platforms, extending the Chinese government's reach into the information space of other countries. Automated bots and algorithms help amplify pro-China narratives and stifle dissent by promoting certain content, creating fake social media profiles and conducting coordinated inauthentic behaviour[62]. China has reportedly used Twitter bots to create and spread disinformation regarding the Hong Kong protests and the COVID-19 pandemic, manipulating narratives on a global scale[63]. There have been instances where state-backed Chinese groups have used Facebook advertising to target users with political messages in many languages, showing a sophisticated understanding of audience segmentation, probably aided by AI tools[64].

---

[56] W. R. Hobbs and M. E. Roberts, 'How sudden censorship can increase access to information', *American Political Science Review*, Vol 112, No 3, 2018, pp. 621-636.

[57] T. Sun and Q. Zhao, 'Delegated censorship: The dynamic, layered, and multistage information control regime in China', *Politics & Society,* Vol 50, No 2, 2022, pp. 191-221.

[58] R. E. Ellis, 'China's Economic Struggle for Position in Latin America', *China Engages Latin America: Distorting Development and Democracy?*, 2022, pp. 85-129.

[59] A. Munoriyarwa, 'The militarization of digital surveillance in post-coup Zimbabwe:' Just don't tell them what we do', *Security Dialogue,* Vol 53, No 5, 2022, pp. 456-474.

[60] B. Buchanan, The hacker and the state: Cyber attacks and the new normal of geopolitics, Harvard University Press, (Cambridge: Untied States), 2020.

[61] Z. Siddiqui, 'Five Eyes intelligence chiefs warn on China's 'theft' of intellectual property', *Reuters*, 18 October 2023.

[62] T. Hsu, and S. L. Myers, 'Pro-China YouTube Network Used A.I. to Malign U.S., Report Finds', *The New York Times*, 14 December 2023.

[63] E. Olcott and S. Yu, 'China escalates zero-Covid propaganda effort as experts warn of economic damage', *The Financial Times*, 14 April 2022; F. King-Wa, 'Propagandization of Relative Gratification: How Chinese State Media Portray the International Pandemic', *Political Communication,* Vol 40, No 6, pp. 788-809, 2023.

[64] S. Lyngaas, 'Meta identifies Chinese propaganda threat ahead of 2024 election', *CNN Business*, 30 November 2023.

China stands as an important example of DPI's employment for state surveillance. Its digital control mechanism, the 'Great Firewall', employs DPI on an unprecedented scale: One of its primary uses is to **monitor discussions** on subjects that the state deems sensitive[65]. Terms related to the 'Tibet autonomy movement', 'spiritual group Falun Gong', 'Tiananmen Square protests', or even specific political figures can trigger the system. Once detected, these communications can be blocked, logged, or even used as a basis for further action against the individuals involved.

China's application of DPI is not limited just to keyword detection. It is a part of a broader strategy to shape the digital narrative and **enforce digital boundaries**[66]. Websites that are perceived as carrying dissenting opinions or alternative narratives are often blocked. DPI aids this by detecting and subsequently preventing access to such content. Additionally, by gauging the topics and sentiments of public digital conversations, Chinese authorities can create **feedback look** and adjust state media propaganda, ensuring that it addresses and shapes public perception effectively.

There have been reports of **monitoring and influence** operations targeting **Chinese nationals** abroad, especially students, through digital means[67]. AI-enhanced surveillance is utilised to track and sometimes intimidate these individuals, often compelling self-censorship or reporting back to the authorities in China. The Chinese government's **United Front Work Department** is known to monitor Chinese diaspora communities and there are reports that it uses digital tools to keep tabs on Chinese nationals abroad and influence their activities[68].

China actively participates in and sometimes seeks to **influence** the formation of **global cyber norms and standards,** including those related to AI. Its goal is to shape these norms to be more permissive of sovereignty-based internet governance, which includes stringent state control over digital spaces. China actively participates in the International Telecommunication Union, a UN body responsible for global telecom networks and standards, where it pushes for international standards that reflect its own domestic policies, which could affect the global governance of AI technologies[69]. As part of its Belt and Road Initiative, China promotes the Digital Silk Road, which includes the export of digital infrastructure that can embed Chinese standards and technologies, including AI, in participating countries.

3.3 Russian algorithmic authoritarianism: The Yarovaya Law of 2016 and evolution after the 2022 Ukraine invasion

Russian state use of digital restrictions as well as AI-based manipulation and election interference methods have been widely debated since the United States (US) elections in 2016. During that year, Russia also passed the **'Yarovaya Law', a package of anti-terrorism laws** that form a coherent, singular legal approach to internet surveillance and censorship[70]. These laws have been seen as a tool to suppress dissent

[65] F. Yang, 'The tale of deep packet inspection in China: Mind the gap', *2015 3rd International Conference on Information and Communication Technology,* IEEE, 2015. pp. 348-351.
[66] R. Sun, L. Shi, C. Yin and J. Wang, 'An improved method in deep packet inspection based on regular expression', *The Journal of Supercomputing*, Vol 75, 2019, pp. 3317-3333.
[67] F. Martin, 'Overseas study as zone of suspension: Chinese students re-negotiating youth, gender, and intimacy. Journal of Intercultural Studies', Vol 39, No 6, 2018, pp. 688-703.
[68] K. Chan and C. Alden, '<Redirecting> the diaspora: China's united front work and the hyperlink networks of diasporic Chinese websites in cyberspace', *Political Research Exchange*, Vol 5, No 1, 2023.
[69] J. Hillman, D. Sacks, J. J. Lew, and G. Roughead, 'China's Belt and Road: Implications for the United States', *New York: Council on Foreign Relations*, 2021.
[70] E. Moyakine and A. Tabachnik, 'Struggling to strike the right balance between interests at stake: The 'Yarovaya','Fake news' and 'Disrespect'laws as examples of ill-conceived legislation in the age of modern technology', *Computer Law & Security Review*, Vol 40, 2021.

and control the narrative within Russia, targeting platforms, services and even individual users who may be seen as threats to the state's narrative.

While the Yarovaya Law by itself is **not directly related to the use of AI,** its data retention requirements both for domestic data collectors as well as foreign technology and social media companies serve as the infrastructure of its other AI-driven surveillance and monitoring projects. One of the most significant aspects of the Yarovaya Law is a **requirement for telecommunications operators and ISPs** to store the content of voice communications, data, images and text messages for six months. This encompasses a vast amount of personal and private information. Furthermore, the law mandates that service providers store metadata (information about the time, location and recipients of communications) for an extended period of three years. This requirement allows for **comprehensive tracking and analysis of communication patterns** without necessarily accessing the content. The law also requires communication service providers to assist security agencies in decrypting any message sent through their networks. This provision potentially undermines encryption and privacy protections, giving security agencies the ability to access private communications. The combination of these highly invasive data collection practices provides the Russian state with a **vast array of personal data formats, such as biometric, text and social media** that can be trained to optimise machine learning classifiers and although the law in itself does not constitute 'algorithmic authoritarianism', it is one of the legal frameworks that enable such repression through mere data collection and storage practices[71].

In a move indicative of Russia's growing emphasis on digital sovereignty and control, the Yarovaya Law was introduced along with other pieces of similar legislation. The state presented a rationale for these stringent measures, framing them within the context of national security imperatives, counter-terrorism objectives and as protective shields against potential foreign interference. This reasoning resonated strongly within the Duma, Russia's parliament, leading to an overwhelming endorsement of these laws. Upon enactment, ISPs and telecommunication operators found themselves bound by new obligations[72]. Not only were they required to store users' data and communications for a designated period, but they were also mandated to grant access to this stored data to security agencies – all without necessitating any court orders. This legislation, **while framed as a protective measure, effectively expanded the state's surveillance capabilities within the digital domain.**

As Russia fortified its digital surveillance architecture, **telecommunication companies and ISPs found** themselves operating in a radically altered landscape. **No longer could they function as neutral entities;** instead, they were now mandated to maintain detailed records of user communications[73]. This was not limited merely to meta-data; the content of communications was also to be archived for a set duration. These rigorous data storage demands coincided with mounting pressure on messaging platforms operating within Russia. Telegram, for instance, found itself in a tight spot as the Russian government exerted pressure on the company to surrender its decryption keys, a move that would undermine the privacy assurances that such platforms extend to their users.

However, perhaps the most audacious leap in surveillance capability was the government's adoption of **Deep Packet Inspection (DPI)** technology. With DPI in play, authorities could not just monitor, but also filter internet traffic in real-time, giving them unparalleled oversight into online content and the power to censor at will[74]. DPI makes possible the detailed inspection and analysis of internet traffic, which when

---

[71] J. W. Lamoreaux and L. Flake, 'The Russian Orthodox Church, the Kremlin, and religious (il) liberalism in Russia', *Palgrave Communications*, Vol 4, No 1, 2018.

[72] K. Ermoshina, B. Loveluck, and F. Musiani, 'A market of black boxes: The political economy of Internet surveillance and censorship in Russia', *Journal of Information Technology & Politics,* Vol 19, No 1, 2022, pp. 18-33.

[73] A. Gurkov, 'Personal Data Protection in Russia', *The Palgrave Handbook of Digital Russia* Studies, 2021, pp. 95-113.

[74] L. Topor and A. Tabachnik, 'Russian Cyber Information Warfare: International Distribution and Domestic Control', *Journal of Advanced Military Studies*, Vol 12, No 1, 2021, pp. 112-127.

combined with AI can enable governments to conduct more sophisticated and extensive surveillance of online activities. AI can process and analyse the vast amounts of data collected via DPI at a scale and speed unattainable by human operators. DPI can also be used to filter and block specific types of content. AI enhances this capability by automatically identifying and censoring content that the government wishes to suppress, such as political dissent, social unrest, or information deemed harmful to national security. The combination of DPI and AI enables the analysis of internet usage patterns, helping to profile individuals' behaviours and interests. This information can be used by authoritarian regimes to identify and target political dissidents, activists and minority groups.

Within Russia, an increase in digital surveillance measures has prompted significant opposition. Key stakeholders, including ISPs and digital corporations, have **expressed concerns** over the new regulations. Moreover, **younger internet** users who are more digitally engaged, have voiced their apprehension about the tightening digital sphere. On the **international front, Western nations and human rights entities** have also raised concerns about Russia's digital policies. All posited that such measures infringed digital rights and freedoms, which further soured Russia's diplomatic interactions with Western countries.

In response, Russian government bodies articulated their position, **emphasising the primacy of national security concerns.** They argued that the measures were analogous to regulations observed in various Western nations. By drawing these parallels, Russian officials aimed at spotlighting inconsistencies perceived in criticisms directed towards their policies. The War in Ukraine intensified this rift, especially the second phase which began in February 2022.

3.3.1 Evolution and current state after Russia's 2022 war on Ukraine

Over time, Russia's legislative landscape concerning the digital realm has grown progressively more stringent. The Yarovaya Law turned out to be **just the start of many regulatory enactments.** Newer legislation has emerged since, targeting what the government has deemed as 'fake news' and online content disrespecting state institutions, further tightening the government's grip over online narratives. Simultaneously, the rapid evolution of AI and machine learning technologies has presented a potential avenue for governments, including Russia's, to augment their surveillance capabilities. Furthermore, there is growing apprehension that Russia might harness these cutting-edge technologies to refine its censorship mechanisms and intensify its monitoring still further.

Amid this tightening regulatory environment, various international technology companies find themselves at a crossroads. Confronted with government directives that could potentially breach user confidentiality, a dilemma is being faced: **acquiesce to demands and potentially jeopardise user trust; or withdraw operations from the Russian** market altogether[75]. According to Scott Marcus et al 'Russia is highly reliant on imports of high-tech goods, with imports worth around USD 19 billion annually. The largest share (45 %) comes from the EU, with 21 % from the US, 11 % from China and 2 % from the United Kingdom. The main import categories are aerospace goods (worth almost USD 6 billion) together with information and communication (nearly USD 4 billion in 2019)'[76]. In recent years, the share of East Asian economies in Russia's total goods imports has risen substantially, with China accounting for most. This trend is particularly evident in technology products such as machinery, equipment and related parts. For instance, in 2019 East Asian countries accounted for about 40 % of Russian machinery and vehicles imports as well

[75] U. A. Mejias and N. E. Vokuev, 'Disinformation and the media: the case of Russia and Ukraine', *Media, culture & society*, Vol 39, No 7, 2017, pp. 1027-1042. H. A. Ünver, 'The Logic of Secrecy: Digital Surveillance in Turkey and Russia', *Turkish Policy Quarterly*, Vol 17, No 2, 2018, pp. 93-103.

[76] J. S. Marcus, N. Poitiers, M. De Ridder and P. Weil, 'The decoupling of Russia: high-tech goods and components', *Bruegel*, 28 March 2022.

as a significant 67 % of imported electrical equipment, with China being the dominant partner in most technology product categories[77].

Since the full-scale invasion of Ukraine in February 2022, Russia's strategy of digital control has intensified and become more technologically intricate. The state's ability to peer into the digital lives of Russian citizens has been bolstered by the deployment of DPI technology across its networks[78]. With the **Sovereign Internet Law**'s evolution, colloquially known as **Runet,** Russia has taken a significant stride in its quest for digital autonomy[79]. This legislation is designed to isolate the Russian internet from the rest of the world, theoretically enabling the Kremlin to maintain an information blackout if deemed necessary. This Runet law essentially gives the Russian government power to centralise control over the internet within Russia. This includes an ability to manage the flow of information and internet traffic, potentially leading to greater control over online content and activities. In extreme scenarios, the government is allowed to isolate Russia's internet segment from the rest of the world. This isolation can facilitate the implementation of AI-driven surveillance tools tailored to the government's needs without external scrutiny or influence[80].

The **Russian government's content-blocking mechanisms** have grown smarter and more pre-emptive. Algorithms have become more pronounced in their deployment and analytics decisions, to the extent that they are now capable of sifting through the internet, thereby stifling dissent before it gains traction. Russian ISPs find themselves in an increasingly tightening grip of regulations, forced to install government-sanctioned filtering hardware that entrenches state surveillance. Moscow's public security cameras, armed with this technology, have reportedly been used to single out and detain protesters, leveraging a network that turns public spaces into surveillance markets[81].

Social media is increasingly becoming a **testing ground for AI-based tools** to monitor and analyse posts and content for signs of subversion. For instance, following Russia's 2022 invasion of Ukraine, there were numerous reports of individuals being targeted for online activities that criticised the country's actions. State surveillance apparatus was probably being used with its sophisticated social media monitoring tools to identify dissenting voices[82].

The sphere of propaganda has witnessed perhaps **the most innovative, albeit insidious, applications of AI.** The Russian government's use of advanced language processing technology has led to more personalised and realistic messages on social media, using bots to promote official narratives[83]. While there is no confirmed use of deepfake technology by the government, it certainly has the potential to create fake videos or audio that could wrongly implicate those who oppose the government. During the Wagner Group rebellion on 23-24 June 2023, several deepfakes emerged on social media, boosted through automated bots, depicting a fake Putin 'surrendering' to Wagner Group chairman Yevgeny Prigozhin[84].

---

[77] J. S. Marcus, N. Poitiers, M. De Ridder and P. Weil, 'The decoupling of Russia', *Bruegel*, 28 March 2022.

[78] Y. Golovchenko, 'Fighting propaganda with censorship: A study of the Ukrainian ban on Russian social media', *The Journal of Politics*, Vol 84, No 2, 2022, pp. 639-654.

[79] R. Ramesh, R. S. Raman, A. Virkud, A. Dirksen, A. Huremagic, D. Fifield, and R. Ensafi, 'Network responses to Russia's invasion of Ukraine in 2022: a cautionary tale for internet freedom', *32nd USENIX Security Symposium*, USENIX Security 23, 2023, pp. 2581-2598.

[80] J. P. Nikkarila and M. Ristolainen, 'RuNet 2020'-Deploying traditional elements of combat power in cyberspace?' in *2017 International Conference on Military Communications and Information Systems*, pp. 1-8, 15-16 May 2017.

[81] S. Hogue, 'Civilian Surveillance in the War in Ukraine: Mobilizing the Agency of the Observers of War', *Surveillance & Society*, Vol 21, No 1, 2023, pp. 108-112.

[82] S. Petrella, C. Miller and B. Cooper, 'Russia's artificial intelligence strategy: the role of state-owned firms', *Orbis*, Vol *65*, No 1, 2021, pp. 75-100.

[83] F. Sufi, 'Social Media Analytics on Russia–Ukraine Cyber War with Natural Language Processing: Perspectives and Challenges', *Information*, Vol 14, No 9, 2023, pp. 485.

[84] J. Jones, 'Deepfake of purported Putin declaring martial law fits disturbing pattern', *MSNBC*, 7 June 2023

This kind of automated information warfare could be a powerful tool for spreading propaganda, further intensifying emergencies and tensions such as the Prigozhin rebellion.

Following its invasion of Ukraine in February 2022, **Russia has significantly ramped up its use of AI-based tools for domestic repression and surveillance,** marking a notable shift in the landscape of state control and monitoring. This intensification reflects a response to the heightened political and social tensions within the country, as well as a need to manage domestic and international perceptions of the conflict. One of the most prominent changes has been **the escalated use of AI in monitoring and censoring digital communication.** The government has employed sophisticated AI algorithms to scrutinise social media platforms and messaging apps. This technology is designed to detect and suppress content automatically that criticises the government's actions in Ukraine or contradicts the official state narrative[85]. In November, Russia further broadened the scope of its AI-based surveillance mechanisms by initiating the 'Oculus Project', which harnesses AI-based text-detection techniques to suppress and censor information related to the Ukraine war, or LGBTQ+ content online[86]. The **Oculus Project** represents a significant advancement in the Russian government's approach to information control. This system, leveraging advanced machine learning and natural language processing, can analyse real-time online data and is not just limited to identifying and censoring content; the system is also designed to track the digital footprints of individuals who disseminate prohibited information, thereby aiding in identifying and potentially prosecuting dissenters. This marks a concerning escalation in digital surveillance capabilities, reflecting a growing trend among authoritarian regimes to utilise cutting-edge technology for internal security purposes.

Moreover, the **surveillance of public spaces** has seen a marked increase, with **facial recognition technology** becoming more prevalent. This is used not only to monitor public areas but also to identify individuals participating in **protests** or **expressing dissenting views.** Journalists, activists and political dissidents in particular have found themselves targeted, with AI-driven surveillance tools being used to track their movements and activities[87]. The government's efforts in AI-driven propaganda have also intensified. Algorithms are now more extensively used to spread pro-government narratives and manipulate public opinion about the war. This digital manipulation extends beyond simple content creation, involving sophisticated strategies to influence public discourse and perception, both domestically and internationally[88].

The role of AI in **cyber warfare** has also evolved. Russia has increased its deployment of AI-driven cyber-attacks, targeting not only Ukrainian infrastructure and communications but also entities in countries perceived as adversaries due to their stance on the conflict[89]. In the face of international sanctions and isolation, the Russian government has **turned increasingly towards domestic technology companies to bolster its AI capabilities.** This collaboration has focused on developing tools for state surveillance and propaganda, reflecting a shift towards self-reliance in technology due to restricted access to global markets and expertise. In late November 2023, President Putin announced a new AI-based initiative that would

[85] L. Masri, 'Facial recognition is helping Putin curb dissent with the aid of U.S. tech', *Reuters*, 28 March 2023.
[86] J. Vainilavičius, 'Russia launches "Oculus" tool to monitor banned information online', *CyberNews*, 15 November 2023.
[87] M. Borak, 'Inside Safe City, Moscow's AI Surveillance Dystopia', *Wired*, 6 February 2023.
[88] O. Robinson, A. Robinson and S. Sardarizadeh, 'Ukraine war: How TikTok fakes pushed Russian lies to millions', *BBC News*, 15 December 2023.
[89] R. Gallagher, 'Russia-Linked Hackers Claim Credit for OpenAI Outage This Week', *Bloomberg*, 9 November 2023.

fundamentally reduce Russian dependence on Western high-technology exports and create a 'self-sufficient' AI ecosystem based in Russia[90].

3.3.2 Russian state actors and algorithmic authoritarianism practices

The **Kremlin**'s approach to algorithmic authoritarianism is an intricate part of Russia's larger strategy not only to maintain power and control within its sovereign territory but also to project its influence abroad. This had been the case since the first invasion of Ukraine in 2014, but a more draconian character was assumed after the main invasion in 2022[91].

Domestically, the Kremlin's interest in AI can be linked to its desire to bolster internal security and maintain social stability to suppress dissent and police non-compliance with the national security measures taken since 2014. **AI technologies are being deployed to monitor and analyse the biometric and digital data** generated by Russian citizens daily, both online and offline. These systems can track digital footprints, predict protests and identify patterns that might signify oppositional behaviour. A particularly relevant example is the use of facial recognition technology in urban surveillance.

Moscow is one of the most heavily monitored and observed cities globally, where cameras equipped with AI can track individuals in real time, making it harder for dissent to manifest in public spaces[92]. The sophistication of AI in these systems facilitates the cross-referencing of visual data with existing databases to flag **individuals who may be of interest to security services.** On the digital front, Russia has reportedly been **using algorithmic techniques to manipulate online information.** Perhaps these capabilities have been studied more within the context of Russia's foreign manipulation and election interference efforts. However, equally potent applications of these techniques are commonplace domestically. **AI-powered bots and trolls** have been implicated in influencing public opinion on social media platforms, both domestically and internationally. These automated systems can flood the digital ecosystem with state-endorsed narratives, drown out dissenting voices, and spread disinformation to create confusion and distrust within communities perceived as adversarial to Kremlin interests[93].

Moreover, the Kremlin's legislative actions underscore its strategy for digital control. Laws that require the localisation of data storage for Russian users arm the state with the potential to access personal information readily, adding yet another layer to their surveillance capabilities. Such laws also facilitate the application of AI algorithms to sift through and analyse this data for any signs of anti-Kremlin sentiments or activities. The **use of AI for censorship is another aspect** of the Kremlin's digital strategy. Increasingly sophisticated algorithms are demonstrating a growing ability to detect and block content that is deemed to challenge state authority or contravene Russian laws, which often include **broad definitions of what constitutes extremist or undesirable content**. Internationally, the Kremlin has been accused of using AI to carry out cyber-espionage activities, attempting to infiltrate the digital infrastructures of other nations to extract information and exert influence[94]. The AI-enhanced capabilities of state-sponsored actors allow them to execute complex cyberattacks and disinformation campaigns with increased efficiency and deniability.

---

[90] L. Varanasi, 'Putin says Russia will develop new AI technology to counter the Western monopoly, which he fears could lead to a 'digital abolition' of Russian culture', *Business Insider*, 26 November 2023.
[91] S. Petrella, C. Miller, and B. Cooper, 'Russia's artificial intelligence strategy: the role of state-owned firms', *Orbis*, Vol 65, No 1, 2021, pp. 75-100.
[92] D. Bazarkina and E. Pashentsev, 'Malicious use of artificial intelligence', *Russia in Global Affairs*, Vol 18, No 4, 2020, pp. 154-177.
[93] R. Thornton and M. Miron, 'Towards the 'third revolution in military affairs' the Russian military's use of AI-enabled cyber warfare', *The RUSI Journal,* Vol 165, No 3, 2020, pp. 12-21.
[94] Z. Thomas, 'Banks' Use of A.I. Raises Risk of Cyberattacks by Russia, Experts Say', *Wall Street Journal*, 23 March 2022.

The **Federal Security Service** (FSB) and the **Main Intelligence Directorate** (GRU) are critical pillars of Russia's security apparatus, employing advanced technologies as a means of bolstering their intelligence and counterintelligence efforts. In the age of information, where data is as critical as physical assets, both agencies have turned to AI to maintain and expand their operational capabilities[95].

FSB, a successor to the KGB, is principally responsible for internal security, counterintelligence and surveillance within Russia. Its adaptation of AI technologies serves various purposes:

- **Monitoring communications:** The FSB utilises AI-driven systems to sift through massive amounts of digital communications. By using natural language processing and machine learning algorithms, the FSB can effectively monitor emails, social media and internet traffic for keywords and patterns that might signal anti-state activities or dissent. Since the second invasion of Ukraine in 2022, Russia has expanded its AI-based monitoring activities in Europe, particularly deploying LLMs in European languages[96].
- **Urban surveillance systems:** Russia has implemented one of the world's most extensive surveillance systems in urban settings. In cities such as Moscow, **AI-driven facial recognition technology** is embedded into the network of CCTV cameras, enabling real-time identification and tracking of individuals[97]. The FSB can harness this technology to follow the movements of suspected dissidents, activists and any other persons of interest.
- **Predictive policing:** By analysing data collected from various sources, AI can predict where and when public disturbances might occur. This predictive analytics capability allows the FSB to deploy resources more effectively to prevent or disrupt gatherings that could challenge state authority[98].
- **Cyber threats:** AI is also a valuable tool in identifying and neutralising cyber threats, with machine learning algorithms that can detect anomalies that suggest a **cybersecurity breach,** thus protecting state interests from both internal and external digital threats.[99]

The GRU, while primarily focused on military intelligence gathering abroad, also leverages AI in certain key areas[100]:

- AI systems can automate the creation and dissemination of disinformation across social media platforms, **targeting foreign populations to influence public opinion** and sow discord, a technique often referred to as '**astroturfing**'[101].
- AI enhances the GRU's capabilities to carry out **cyberattacks** against foreign governments, institutions and infrastructure. These AI systems can make thousands of attempts to breach security in a matter of

[95] V. Akimenko and K. Giles, 'Russia's cyber and information warfare', *Asia Policy*, Vol 15, No 2, 2020, pp. 67-75.
[96] D. Mac Dougall, 'Spies like us: How does Russia's intelligence network operate across Europe?', *EuroNews*, 18 August 2023.
[97] I. Borogan, A. Soldatov, E. Grossfeld and D. Richterova, 'What impact has the war on Ukraine had on Russian security and intelligence?', *King's College London*, 22 February 2023.
[98] V. Gubko, M. Novogonskaya, P. Stepanov and M. Yundina, 'AI And Administration Of Justice In Russia', e-Revue Internationale de Droit Pénal. A-07, 70586803X, 12 April 2023. A. Zharovaa, V. Elin and P. Panfilov, 'Introducing Artificial Intelligence Into Law Enforcement Practice: The Case Of Russia', 30th DAAAM International Symposium On Intelligent Manufacturing And Automation, 2019, pp. 688-692.
[99] A. Soldatov and I. Borogan, 'Russian Cyberwarfare: Unpacking the Kremlin's Capabilities', *Center for European Policy Analysis*, 8 September 2022.
[100] R. Heickero, 'Russia's information warfare capabilities. In Current and Emerging Trends in Cyber Operations: Policy, Strategy and Practice', *Palgrave Studies in Cybercrime and Cybersecurity*, 2015, pp. 65-83.
[101] N. Guggenberger and P. N. Salib, 'From Fake News to Fake Views: New Challenges Posed by ChatGPT-Like AI', Lawfare Institute, *The Brookings Institution*, 20 January 2023.

minutes, far faster than any human could, particularly reinforcing **distributed denial-of-service (DDoS)** attacks that leverage speed and scale to overwhelm digital systems[102].

- AI algorithms can process vast amounts of intercepted communications more quickly and efficiently than human operatives, which is essential for the GRU's Decryption and Signals Intelligence operations. AI may also assist in breaking encrypted messages and analysing patterns in data traffic to **gather intelligence**[103].
- The GRU also invests in AI to enhance **Russia's military capabilities,** including the development of autonomous weapons systems and drones that can be used for reconnaissance or targeted operations[104].

In essence, the FSB and GRU are **using AI as a force multiplier to enhance their traditional intelligence and surveillance tasks.** These technologies enable them to operate with greater stealth, precision and efficiency, representing a significant upgrade in their ability to identify, monitor and neutralise perceived threats to state security. They serve as a conduit between Russia's domestic algorithmic authoritarianism practices and its foreign security applications; the TTPs Russia deploys in foreign security operations feed into its domestic applications and vice versa. To that end, it is not reliably possible to separate external and internal strategies of algorithmic authoritarianism in Russia, as both domains feed into each other regarding the technologies and approaches utilised.

The Federal Service for Supervision of Communications, Information Technology and Mass Media (hereafter **'Roskomnadzor'**) system scans text and images across various online platforms, including social media, fora and news outlets. The AI algorithms are designed to identify content considered 'extremist' or contrary to state policies, using natural language processing to detect nuances that might be missed by human monitors. The AI tools employed by Roskomnadzor can already efficiently filter through texts, images and videos, identifying as well as acting upon sensitive content with a high degree of accuracy. In February 2023, the agency announced plans to expand the use of AI still further for monitoring manipulation and polarisation online[105]. These systems, equipped with a database of sensitive keywords, can perform real-time censorship, autonomously blocking content before it becomes widely accessible. This mechanism allows Roskomnadzor to maintain strict control over the information disseminated across the Russian internet[106].

Roskomnadzor's AI tools have a **dual function** beyond just censoring content. They are used not only to **suppress information** that deviates from the state's approved narrative but also to **monitor public sentiment by analysing expressions** across various platforms. Recently, the agency declared that it had uncovered various AI botnets (robotic networks) that have collected web data for foreign governments and boosted its efforts to detect text and content automatically, which is aimed at harvesting Russian user

[102] G. Wilde, 'Cyber Operations in Ukraine: Russia's Unmet Expectations', *Carnegie Endowment for International Peace*, 12 December 2022.
[103] Insikt Group, 'Obfuscation and AI Content in the Russian Influence Network "Doppelgänger" Signals Evolving Tactics', *Recorded Future – Russia Threat Analysis*, 5 December 2023.
[104] B. Laird, 'The Risks of Autonomous Weapons Systems for Crisis Stability and Conflict Escalation in Future U.S.-Russia Confrontations', *RAND Corporation*, 3 June 2020.
[105] Novaya Gazeta Europe, 'Roskomnadzor plans to use AI to monitor 'manipulations and social polarisation' online', *Novaya Gazeta Europe*, 8 February 2023.
[106] E. Gaufman, 'Cybercrime and Punishment: Security, Information War, and the Future of Runet', in D. Gritsenko, M. Wijermas and M. Kopotev (eds), *The Palgrave Handbook of Digital Russia Studies,* 2021, pp. 115-134.

data[107]. This gives the state justification not just for removing content, but also for subtly influencing public opinion by alerting citizens about 'automated foreign influence'.

Additionally, AI acts as a **strict enforcer of compliance**, **ensuring adherence to rules related to information technology and telecommunications.** These systems can detect activities such as the unauthorised use of VPNs and efforts to circumvent digital restrictions imposed by the state[108]. The use of AI by Roskomnadzor marks a significant advancement in the state's ability to oversee and influence the information environment. By leveraging AI for state policy, the government strengthens its control over digital spaces, blending traditional governance methods with modern technology. The integration of AI into informational strategies indicates a sophisticated approach to managing digital narratives and policing digital public debate, representing a significant expansion of state authority into the digital realm[109].

3.4 Iranian AI-based repression systems: Silencing dissent and suppressing opposition

Although Iranian algorithmic authoritarianism practices are technically behind those of Russia and China, Iran's **NIN (National Information Network)** is a significant attempt to isolate Iranian users from the global Internet, by using several AI-based surveillance and monitoring tools that constitute algorithmic authoritarianism cases[110]. Conceived in the early 2010s, NIN's foundational logic has its roots in the political upheavals of 2009. It was a time when the Iranian population, increasingly digitally interconnected, started voicing political dissent. This digital momentum then yielded physical results in the form of widespread protests and riots across Iran. To curtail this digital revolution and regain the narrative, the state conceptualised NIN as a means of diminishing Iran's reliance on the global internet. This network, far more than just infrastructure, represents Iran's larger objectives to create a broader technological communication base that will allow the state to develop advanced technologies such as AI, independent of information flows from the so-called West[111]. On the one hand, it aimed at **reclaiming digital autonomy**, ensuring that the content consumed by its citizens aligns with the state's ideologies and is not influenced by what is perceived as a Western 'cultural invasion'[112]. On the other hand, it **seeks to fortify the state's surveillance apparatus,** giving it total control over its population's online activities without the need to deal with Western service providers or social media companies[113]. This establishes the data infrastructure for Iran's algorithmic authoritarianism practices, as the data needed to feed in Iran's comparatively moderate AI-based tools (compared with China and Russia) can be collected at scale without dealing with foreign stakeholders – such as social media platforms or technology companies.

A defining characteristic of NIN is its emphasis on domestically hosted content. By prioritising **local hosting,** Iran ensures swift access for its users (public justification for NIN) while retaining firm control over content moderation. This control extends to the promotion of Iranian-made digital services, search engines, email providers and even social media platforms, seeking to rival their global counterparts.

---

[107] TASS Russian News Agency, 'Russian Internet regulator uncovers AI bot gathering web data for foreign machine learning', 13 December 2023.
[108] 'Reuters, 'Russia plans to try to block VPN services in 2024 – senator', 3 October 2023.
[109] N. Maréchal, 'Networked authoritarianism and the geopolitics of information: Understanding Russian Internet policy', *Media and communication*, Vol 5, No 1, 2017, pp. 29-41.
[110] L. Namdarian, S. Alidousti, and B. Rasuli, 'Developing a comprehensive framework for analyzing national scientific and technical information policy: application of HeLICAM in Iran', *Online Information Review*, Vol 45, No 7, 2021, pp. 1381-1403.
[111] R. Stone, 'Iran's researchers increasingly isolated as government prepares to wall off internet', *Science*, 11 September 2023.
[112] J. B. Alterman, 'Protest, Social Media, and Censorship in Iran', *Center for Strategic and International Studies*, 18 October 2022.
[113] A. Akbari and R. Gabdulhakov, 'Platform surveillance and resistance in Iran and Russia: The case of Telegram', *Surveillance & Society*, Vol 17, No 1 and 2, 2019, pp. 223-231.

Internet nationalisation and digital services in Iran create a drive towards indigenous systems and software, with the benefit of boosting national research and development by creating significant demand for systems that could otherwise be imported from the West. By creating clones of these systems and maintaining them indigenously, Iran also creates significant demand for domestic engineering, boosting national 'Science, Technology, Engineering and Mathematics' (or STEM) disciplines and serving as another layer of justification for NIN.[114] However, while this strategy champions Iranian digital products, it also acts as a **gateway for state-led censorship, since all of the systems and software – as well as their developers – are in Iran,** and can be controlled by the government at will.

The same indigenisation can be seen in news production and information flows. Various **foreign websites,** especially those that could provide counter-narratives, find themselves **frequently blocked,** pushing the populace to gravitate toward domestic alternatives. This mechanism is further strengthened by state-controlled internet gateways, which have the dual capability of directing all of Iran's internet traffic and when necessary, shutting down access to the global internet, leaving only the NIN as functional[115]. The existing NIN infrastructure presents an **ideal precursor for more advanced algorithmic authoritarianism practices in Iran** and ensures that regardless of how these technologies evolve, NIN serves as the foundational data infrastructure of future AI-based systems.

Beyond the realm of digital access and control, NIN also signals Iran's **pursuit of 'digital sovereignty'.** In the Iranian state narrative, this self-reliance in the digital domain serves as a shield, 'guarding the nation against potential cyber threats', external digital interferences and even international sanctions targeting its internet infrastructure. This narrative is intended to **stifle public opposition against digital isolation** and discursively construct it as a 'necessary precaution', thereby stoking nationalism. Activists, journalists and citizens tread cautiously, perpetually aware of **omnipresent surveillance**[116], and the nationalist narrative around such practices of repression ensures the domestic weakness of criticisms.

The integration of AI-based surveillance technologies into Iran's governance framework has also undergone a systematic evolution. As the country's urban infrastructures incorporate high-definition surveillance equipment, the state's surveillance apparatus has expanded its reach into public spaces. These **cameras, equipped with facial recognition capabilities,** have been strategically positioned to capture significant public activity. In recent years, there have been **several partnership agreements between China and Iran aimed** at bolstering Iran's surveillance and AI-based repression practices[117]. Such deployment was particularly evident during periods of civil unrest, such as the **Iranian protests of 2019–2020,** when these **technologies played a role beyond passive observation,** facilitating rapid identification and, in some cases, subsequent detention of individuals based on facial data metrics[118]. These data mining tools systematically accumulate information, constructing comprehensive digital profiles by analysing patterns in social media engagement, content sharing and networking behaviour patterns. Most recently, these AI-based tools were deployed against women's rights movements in Iran[119]. This data-driven

[114] UN Conference on Trade and Development (UNCTAD), 'Science, Technology and Innovation Policy Review – Islamic Republic of Iran', UNCTAD, 2016.
[115] C. Adebahr and B. Mittelhammer, 'Upholding Internet Freedom as Part of the EU's Iran Policy', *Carnegie Europe*, 29 November 2023.
[116] M. Michaelsen, 'Far away, so close: Transnational activism, digital surveillance and authoritarian control in Iran', *Surveillance & Society*, Vol 15, Issue 3 and 4, 2017, pp. 465-470.
[117] T. Ryan-Mosley, 'This huge Chinese company is selling video surveillance systems to Iran', *MIT Technology Review*, 15 December 2021; J. Askew, 'China turbocharging crackdown on Iranian women, say experts', *EuroNews*, 14 April 2023.
[118] A. Shahi and E. Abdoh-Tabrizi, 'Iran's 2019–2020 demonstrations: the changing dynamics of political protests in Iran', *Asian Affairs*, Vol 51, No 1, 2020, pp. 1-41.
[119] R. George, 'The AI Assault on Women: What Iran's Tech Enabled Morality Laws Indicate for Women's Rights Movements', *Council on Foreign Relations*, 7 December 2023.

approach to monitoring means that every digital interaction, from social media posts to networking affiliations, contributes to an individual's digital record, significantly compromising online anonymity and ensuring sustained algorithmic authoritarianism at scale.

Sources of **Iran's total high-technology imports** in recent years have been China, Turkey, India and the United Arab Emirates (UAE). Broadcasting equipment (surveillance cameras, radio transmitters, computer hardware) and computers constitute the largest share of Iran's high-technology imports (13.4 % in the latest recorded year of 2021) and 98.3 % of these components are exported from the UAE[120]. Although China is a major player in Iran's surveillance and monitoring projects and initiatives, most of its hardware infrastructure comes from the UAE[121].

A hallmark of Iran's AI-based approaches to steering public debate is the augmentation of social media platforms with **AI-driven bots and automated accounts**[122]. These bots are programmed to post, share, comment and amplify content that supports the regime's perspective, thereby flooding the information space to drown critical debate and discussion. Unlike traditional propaganda mechanisms, these bots can operate at a staggering pace and scale. By analysing trending topics and popular hashtags, they can swiftly inject state-favoured narratives into mainstream digital discourse. This approach not only amplifies the reach of propaganda but also lends an illusion of organic support and consensus, given the volume of engagement such content garners. Simultaneously, Iran employs sophisticated algorithms to monitor and analyse digital content, which aids in swiftly identifying dissenting voices or narratives. Once identified, the AI systems prioritise the promotion of counter-narratives. Most social media platforms have been criticised for being slow or apathetic towards Iranian bots that are active on their systems[123].

For instance, during periods of civil unrest or international disputes, there is a noticeable surge in online content that underscores state perspectives, often drowning out opposing viewpoints. Moreover, the Iranian government taps into **natural language processing (NLP)** tools for sentiment analysis[124]. By analysing sentiment within vast swathes of digital content, the regime can gauge public opinion on specific issues, policies and events. This real-time feedback loop allows the authorities to refine and adapt their propaganda strategies with impressive agility. The content generation process itself has been influenced by AI. **Deep learning models have been employed to craft content,** ranging from written articles to more visually appealing infographics and video snippets, which resonate with specific demographics. Such content when paired with **micro-targeting algorithms**[125] can be **directed at specific segments of the population,** ensuring maximum impact. Internationally the Iranian government utilises **similar AI-driven tactics to shape narratives,** especially in languages other than Persian. Automated

[120] Observatory of Economic Complexity, 'Economic Complexity Indicators of the Islamic Republic of Iran', 2021-2022.

[121] S. Kerr and N. Bozorgmehr, 'UAE boosts trade with Iran after eased restrictions on business activity', *Financial Times*, 10 September 2023; B. Faucon, 'U.A.E. Trade Provides Iran With Western Goods, From Perfume to Laptops', *The Wall Street Journal*, 5 July 2022.

[122] M. Elswah and M. Alimardani, 'Propaganda Chimera: Unpacking the Iranian Perception Information Operations in the Arab World', *Open Information Science*, Vol 5, No 1, 2021, pp. 163-174.

[123] L. H. Newman, 'Instagram Slow to Tackle Bots Targeting Iranian Women's Groups', *Wired*, 19 July 2022.

[124] M. Grinko, S. Qalandar, D. Randall, and V. Wulf, 'Nationalizing the Internet to Break a Protest Movement: Internet Shutdown and Counter-Appropriation in Iran of Late 2019', *Proceedings of the ACM on Human-Computer Interaction*, Vol 6, No CSCW2, 2022, pp. 1-21.

[125] Microtargeting is a marketing and consumer research technique that uses consumer data and demographics to identify the interests of potential buyers and steer purchasing preferences. This technique was deployed in the Cambridge Analytica scandal and is used by most governments to monitor electoral preferences and has implications in political science: F. Votta, 'Algorithmic Microtargeting? Testing the Influence of the Meta Ad Delivery Algorithm', European Consortium for Political Research, Joint Sessions of Workshops, Sciences Po Toulouse, 25-28 April 2023; Also see: European Commission, Study on the impact of new technologies on free and fair elections, DG JUST Election Study, March 2021.

translation tools, enhanced by AI, ensure that the state's perspective is shared across many linguistic platforms, thereby reaching a global audience. This international focus is evident in the numerous English, Arabic and other language-based digital campaigns that emerge during key geopolitical events.

3.4.1 Key actors in Iranian algorithmic authoritarianism

The **Supreme Leader** of Iran, Ayatollah Ali Khamenei, sits at the apex of Iran's political structure and wields considerable influence over national strategy, particularly in areas concerning societal control and censorship. Allied closely with the **Guardian Council,** a body charged with the task of aligning national policies with Islamic jurisprudence, the Supreme Leader's office shapes the overarching goals and strategies for domestic surveillance and control, employing AI as a potent tool for these purposes[126]. In this collaborative governance model, the Supreme Leader's ideological tenets serve as a guiding framework for the nation's internal security policies, focusing heavily on preserving the Islamic Republic's principles and limiting dissenting views. This framework informs the development and deployment of AI-driven technologies aimed at monitoring the population, ensuring that citizens' activities, both online and offline, remain within the boundaries set by the state's religious and moral codes[127].

AI systems, under the directives of the Supreme Leader and the Guardian Council, are instrumental in various areas of algorithmic authoritarianism. They **monitor internet traffic and social media platforms** for signs of **anti-government sentiment or mobilisation,** employing sophisticated algorithms capable of text and image recognition to scan for content that may be considered subversive or non-compliant with the state's ideology[128]. These tools can identify patterns indicative of dissident behaviour, which enable authorities to pre-emptively address and neutralise potential threats to the regime's stability. Furthermore, the Guardian Council's responsibility to vet **political candidates and ensure the Islamic suitability of legislation has extended into the digital realm.** AI tools are utilised to scrutinise the digital footprints of individuals seeking political office, assessing their adherence to Islamic and revolutionary values. Such technology grants the Guardian Council an unprecedented capacity to influence political participation and maintain the theocratic integrity of the state apparatus[129].

At the citizen level, the implementation of AI in **public surveillance** systems has facilitated **real-time monitoring of large crowds,** with facial recognition software capable of singling out individuals for further scrutiny. In urban environments, these technologies contribute to a pervasive state presence, reminding citizens of the authorities' constant vigilance[130]. The use of AI in these capacities reflects a broader strategy by the Supreme Leader and the Guardian Council to retain control over the narrative within Iran. By enforcing strict digital governance and utilising AI to police public discourse, the leadership ensures that its ideological priorities are reflected across all levels of society, thereby reinforcing its grip on

---

[126] M. Eslami, N. S. Mosavi, and M. Can, 'Sino-Iranian cooperation in artificial intelligence: A potential countering against the US Hegemony,' *The Palgrave Handbook of Globalization with Chinese Characteristics: The Case of the Belt and Road Initiative,* 2023, pp. 543-559.

[127] Very recently, a religious debate in Iran focused on whether AI can issue religious rulings (*fatwas*). Iran's religious leadership leans positively towards harnessing the power of AI to scale religious rulings. On how the clergy in Iran see AI, please refer to: N. Bozorgmehr, 'Robots can help issue a fatwa': Iran's clerics look to harness AI', *Financial Times*, 24 September 2023.

[128] C. A. Wege, 'Iranian counterintelligence', *International Journal of Intelligence and Counterintelligence*, Vol 32, No 2, 2019, pp. 272-294.

[129] M. Ghiabi, 'The council of expediency: crisis and statecraft in Iran and beyond', *Middle Eastern Studies*, 2019, Vol 55, No 5, pp. 837-853.

[130] C. Alkhaldi and N. Ebrahim, 'Iran proposes long jail terms, AI surveillance and crackdown on influencers in harsh new hijab law', *CNN Middle East*, 2 August 2023.

the nation's political and social life[131]. This fusion of traditional theocratic rule with cutting-edge technology represents a powerful means of sustaining the regime's authority in an increasingly connected digital world.

One prominent example is the **application of Islamic moral standards to digital content.** Iran's approach to internet censorship is heavily influenced by Islamic jurisprudence, which emphasises the importance of moral and religious propriety. This is evident in the state's use of AI to scale access to social media platforms and online content algorithmically. **Websites or online material that are deemed inconsistent with Islamic values** – such as those promoting political dissent against the theocratic regime or containing content considered morally corrupting – are **frequently blocked.** This effort aligns with the Islamic concept of '*Amr bil Maroof wa Nahi anil Munkar*' (Commanding what is just and forbidding what is evil), a principle deeply ingrained in Iran's theocratic governance. Surveillance technologies are another area where theocratic principles intersect with modern tools. The government's monitoring of digital communications and online activities reflects a broader Islamic principle of maintaining social order and religious adherence. This surveillance is justified within the framework of protecting Islamic values and ensuring public conformity to religious norms, reminiscent of the historical role of 'Mutaween' (religious police) in enforcing moral standards[132].

The Revolutionary Guard Corps (IRGC) is a significant military organisation within Iran, which plays a crucial role in the country's national security apparatus, extending its reach both domestically and internationally. With a distinct cyber division, the IRGC has developed advanced capabilities in digital surveillance and intelligence, leveraging AI as a key component within its strategy[133]. AI systems enable the IRGC to sift through enormous quantities of data on the internet, including social media posts, emails and other forms of digital communication. These systems are trained to detect patterns, keywords and sentiments that may indicate opposition to the Iranian regime. By automating the process of data analysis, the IRGC can efficiently monitor and interpret digital behaviour simultaneously.

Once potential dissidents or opposition groups are identified, the IRGC can utilise this intelligence in coordination with law enforcement and judicial authorities to intervene, detain, or otherwise suppress the flow of information by and across these entities. The **deployment of AI in this manner enhances the IRGC's capacity to maintain a tight hold over public discourse and curb activities that it perceives as threats to national security** or the prevailing political order. Additionally, AI tools have been reportedly used by the IRGC to create and maintain profiles of individuals across various platforms, enabling a more comprehensive surveillance approach that can track a person's online footprint. This digital profiling aids in understanding social networks, personal associations, and the spread of information that could mobilise opposition against the state[134].

A report by **IPVM** (Internet Protocol Video Market), a surveillance research group, reveals that **Tiandy,** a major Chinese video surveillance company, is supplying surveillance technology to various Iranian institutions, including the IRGC, police and military authorities. This technology includes advanced AI features such as facial recognition and race detection, together with devices such as 'smart' interrogation tables paired with 'tiger chairs', known for their use in torture. This situation highlights **China's expanding**

---

[131] T. Saheb, 'Ethically contentious aspects of artificial intelligence surveillance: a social science perspective', *AI and Ethics*, Vol 3, No 2, 2023, pp. 369-379.

[132] An excellent overview of how Iran's theological precursors affect surveillance and censorship, See: B. Rahimi, 'Censorship and the Islamic Republic: Two Modes of Regulatory Measures for Media in Iran', *Middle East Journal*, Vol 69, No 3, pp. 358–78, 2015.

[133] U. Banerjea, 'Revolutionary intelligence: The expanding intelligence role of the Iranian Revolutionary Guard Corps', *Journal of Strategic Security*, Vol 8, No 3, 2015, pp. 93-106.

[134] A. M. Tabatabai, 'Other side of the Iranian coin: Iran's counterterrorism apparatus', *Journal of Strategic Studies*, Vol 41, No 1-2, 2018, pp. 181-207.

strategic relationship with Iran, particularly in the realm of surveillance technology exports to authoritarian regimes. Tiandy's products, especially the controversial 'ethnicity tracking' tool, have been implicated in the Uyghur minority's repression, as mentioned earlier. Tiandy, with significant sales globally, has secured a five-year contract in Iran and plans to establish a local presence. The company, though privately owned, has close ties with the Chinese government and its CEO, Dai Lin, is a known supporter of the Communist Party. The report also raises concerns over **US sanctions violations,** as Tiandy's networked video recorders used by the Iranian military contain chips from US manufacturer Intel. This finding has prompted Intel to initiate an investigation into the matter. The broader context of this report points to Iran's efforts to build a digital control system over its citizens, following China's model and utilising Chinese tools. This includes adopting aspects of China's 'social credit' system and upgrading surveillance infrastructure with Chinese technology, as seen in a past deal with **ZTE,** a Shenzhen-based company[135]. IRGC's camera systems also police hijab violations, using AI-based feature detection techniques[136].

3.4.2 Transnational surveillance and control

Iran's use of AI to engage in international repression and algorithmic propaganda can be outlined across key areas, including cyber operations, social media manipulation and digital surveillance:

- **International hacking campaigns:** Iran has been accused of conducting cyber espionage campaigns that target government officials, activists and businesses around the world. These operations often involve AI-powered tools for phishing, social engineering and malware distribution that can learn and adapt to the defences of different systems. Iran's hacking capabilities have been on display in various incidents, such as the cyber-attacks on American financial institutions and the Bowman Avenue Dam in New York, which were widely attributed to Iranian hackers[137]. The use of AI in these incidents is not documented in public reports, but the complexity of such cyber operations often suggests an advanced level of automation and adaptability that could be enhanced by AI technologies.
- **Social media manipulation:** Similar to China, Iran has been reported to use bot networks on social media platforms to amplify pro-Iranian narratives and spread disinformation. AI algorithms are deployed to create and disseminate content, manage accounts and engage with users to influence public opinion on topics such as the Iranian nuclear programme or its regional policies. Automated accounts, aided by machine learning techniques, spread propaganda that aligns with Iranian state interests. These accounts can be programmed to target specific demographics and regions, adapting content to the cultural contexts of different audience groups to maximise impact. In a significant case study, political scientist Daniel Byman has explored how Tehran's operatives ran fake accounts and pages on Facebook and Twitter, mimicking real news organisations and political groups[138]. These operations employed AI to manage the scale and complexity of maintaining multiple personas and tailoring content to different audiences.
- **Surveillance technologies and the monitoring of dissidents abroad:** Iranian intelligence services monitor the activities of Iranian dissidents living overseas.[139] This has recently begun to involve the use of AI for analysing large datasets gathered from social media, communication intercepts as well as other digital footprints to identify and track targets[140]. The 2019 expulsion of two Iranian diplomats

[135] T. Ryan-Mosley, 'This huge Chinese company is selling video surveillance systems to Iran', *MIT Technology Review*, 15 December 2021

[136] K. Johnson, 'Iran Says Face Recognition Will ID Women Breaking Hijab Laws', *Wired*, 10 January 2023.

[137] J. Berger, 'A Dam, Small and Unsung, Is Caught Up in an Iranian Hacking Case', *The New York Times*, 25 March 2016.

[138] D. Byman, 'The Social Media War in the Middle East', *The Middle East Journal*, Vol 75, No 3, 2021, pp. 449-468.

[139] Human Rights Watch, 'Iran: Targeting of Dual Citizens, Foreigners', 26 September 2018.

[140] B. Foucon, 'Iran Shifts Tactics to Use Covert Police, Tech to Crack Down on Protests', *The Wall Street Journal*, 18 October 2022.

from the Netherlands was linked to the assassination of two Dutch nationals of Iranian origin[141] who were being monitored through high-technology tools deployed by the Iranian cyberespionage group **'Charming Kitten'.** Germany's Federal Office for Protection of the Constitution published a report in August 2023, singling out this group for its reach into Germany and its targeting of Iranian dissidents as well as expatriates using AI[142]. These incidents highlighted the actual reach of Iranian intelligence activities, which are increasingly incorporating AI-based data analysis to identify and monitor targets abroad and engage in information suppression.

- **Censorship and content control and the internet infrastructure:** Iran has engaged in AI-based Foreign Information Manipulation and Interference (FIMI) attempts in Europe to suppress discussions and censor content related to Iran protests in January 2023[143]. These tools were used to disrupt overseas dissidents' ability to communicate and control the information environment by targeting the Iranian Diaspora's VPN provision attempts to their compatriots living in Iran[144]. The role of the Iranian Diaspora in Europe in helping domestic dissidents circumvent censorship efforts by the government is increasingly being targeted by the Iranian government in turn, using AI-based profiling and location-detection tools to intimidate and suppress these efforts[145].

3.5 Egypt or the quest to prevent another Tahrir

In recent years, Egypt has increasingly turned to AI as a tool for political repression and surveillance, both within its borders and against dissidents abroad. This trend represents a consistent evolution in the government's approach to monitoring and controlling dissent since the Tahrir Revolution, presenting an approach that has become more sophisticated and far-reaching with the advent of advanced technology[146].

As far back as **2014,** the Egyptian government had plans to use **AI for surveillance and political repression,** driven particularly by the so-called Arab Spring in 2011. Notably, the Interior Ministry's leaked tender revealed plans for a sophisticated mass surveillance system, designed to monitor platforms such **as Facebook, Twitter, YouTube, WhatsApp, Viber and Instagram systematically**[147]. This system was to scan social media networks for 26 specific topics, ranging from defamation of religion to calls for illegal demonstrations and terrorism, though the full list remains undisclosed. Since then, successive governments have monitored electronic communications, leading to the arrest and prosecution of activists for their social media posts. Thousands of former President Mohamed Morsi's supporters, for instance, have been detained for exercising their rights to freedom of expression and assembly, with reports of peaceful protestors being arrested, tortured and ill-treated. The crackdown extended to individuals posting on platforms such as YouTube, Facebook and Twitter[148]. The surveillance initiative was framed by the Interior

[141] M. Levitt, 'Iran's Deadly Diplomats', *CTC Sentinel*, Vol 16, 2018, pp. 10-15; Reuters Staff, 'Dutch recall ambassador to Iran after diplomats expelled', *Reuters*, 4 March 2029.
[142] Bundesamt für Verfassungsschutz, 'BfV Cyber-Brief Nr. 01/2023: Advisory on cyber espionage against critics of the Iranian regime in Germany', 10 August 2023.
[143] P. Dave, 'Tech Workers Fight for Iran Protesters as Big Tech Plays It Safe', *Wired*, 20 January 2023.
[144] D. Kilic and A. Ni. Chulain, 'How Iranians are hopping between VPNs to stay connected and break through Internet censorship', *EuroNews*, 6 November 2022.
[145] Freedom House, 'Iran Freedom of the Net 2022 Report', 2023.
[146] M. O. Jones, Digital authoritarianism in the Middle East: Deception, disinformation and social media, Hurst Publishers (London: United Kingdom), 2022.
[147] N. Pratt and D. Rezk, 'Securitizing the Muslim Brotherhood: State violence and authoritarianism in Egypt after the Arab Spring', *Security Dialogue*, Vol 50, No 3, 2019, pp. 239-256.
[148] N. Sayed, 'Towards the Egyptian revolution: Activists' perceptions of social media for mobilization', *Journal of Arab & Muslim Media Research*, Vol 4, No 2-3, 2012, 273-298.

Minister as a measure to combat terrorism and protect national security. He indicated that to track certain individuals the system would employ specific search terms related to activities considered illegal under Egyptian law.

The Egyptian government's use of AI for surveillance, especially in monitoring social media and tracking digital communications, has brought to light various real-world examples that demonstrate its impact on civil society and individual freedoms. One notable instance involved the **sentencing of social media influencers Mawada Eladhm and Haneen Hossam, who faced charges of 'human trafficking',** with substantial fines and prison sentences[149]. Despite their content being non-political, the government monitored their activities on platforms such as Instagram, Twitter and TikTok. These cases underscore the extent to which social media monitoring extends beyond traditional political activism to broader aspects of online expression.

The implementation of **Egypt's Anti-Cyber and Information Technology Crimes Law,** ratified in 2018, has given authorities wide-reaching powers. This law aims at combating extremist and terrorist organisations but has also been used to punish actions that are seen as violating 'the values and principles of the family in Egyptian society'[150]. The law allows for the blocking of websites and the monitoring of online content deemed threatening to national security or the economy. The situation for bloggers and social media influencers in Egypt has worsened, with a security campaign specifically targeting female content creators on platforms like **TikTok** and **Likee.** At least ten women have been convicted since the campaign began, illustrating the government's focus on social media as a domain for enforcing societal norms and silencing dissent. The cases of Hossam and Eladhm, who were found guilty under the cybercrime law for 'violating family principles and values in Egyptian society', highlight the broad and subjective application of the law. The charges against them were based on encouraging women to monetise their video clips on Likee, which prosecutors deemed contrary to Egyptian societal values.

The **recent integration of advanced facial recognition technology by Telecom Egypt** will have broader implications beyond its stated purpose of data centre security. In Egypt, where the political landscape is often turbulent, the adoption of facial recognition systems in public or semi-public spaces such as data centres could be perceived as a stepping stone to more expansive monitoring[151]. The efficiency of these systems in identifying individuals almost instantaneously presents a powerful tool for state surveillance. This tool, if or when applied to public surveillance, **could effectively deter and identify participants in political protests,** potentially chilling public dissent. When individuals know that their presence at a protest could be easily and permanently recorded, it may dissuade them from participating in such activities, thereby undermining the right to peaceful assembly. Moreover, the training of staff and the creation of a Competence Center by Audio Technology SAE suggest an investment in the infrastructure necessary to support the widespread deployment and operation of facial recognition systems[152]. This could indicate a future where such surveillance is not an isolated practice but **an integrated part of the public security apparatus.** While the current use of facial recognition by Telecom Egypt is focused on security and operational efficiency within a data centre, the underlying technology has the potential to be repurposed for broader applications[153].

---

[149] Amnesty International, 'Egypt: Women influencers jailed over 'indecency': Hanin Hossam, Mawada el-Adham', 14 July 2021.
[150] A. M. Abozaid, 'Digital Baltaga: How Cyber Technology Has Consolidated Authoritarianism in Egypt', *SAIS Review of International Affairs*, Vol 42, No 2, 2022, pp. 5-19.
[151] M. S. AlAshry, 'A critical assessment of the impact of Egyptian laws on information access and dissemination by journalists', *Cogent Arts & Humanities*, Vol 9, No 1, 2022.
[152] B. Hassib and J. Shires, 'Manipulating uncertainty: cybersecurity politics in Egypt', *Journal of Cybersecurity*, Vol 7, No 1, 2021.
[153] Privacy International, 'State of Privacy in Egypt 2019', 26 January 2019.

The **reach of Egypt's AI-driven surveillance goes beyond its territory,** targeting Egyptian expatriates and dissidents living abroad, for which the government has reportedly used sophisticated cyber-espionage tools, including AI[154]. This global surveillance network **suggests a collaboration with other governments and international private technological firms,** although the full extent of these partnerships often remains shrouded in secrecy. In October 2019, a report detailed that the Egyptian government had engaged in cyber espionage activities targeting Egyptian dissidents, which included installing mobile applications on the targets' phones to extract files, track locations and identify contacts[155]. The victims of these surveillance activities were identified as Egyptian journalists, academics, lawyers, opposition politicians and human rights activists. This government action falls under the heading of espionage and deliberately targeted civil society, suggesting a systematic approach to monitoring and potentially repressing dissent both within and possibly outside of Egypt's borders.

The use of AI-based surveillance and monitoring technologies in Egypt is reflective of a global trend where **such tools are increasingly harnessed by governments for various reasons, including public security**. In the case of Egypt, concerns have been raised by the international community regarding the potential misuse of these technologies for political repression. Digital rights groups have drawn attention to the deployment of a vast network of CCTV cameras in **Egypt's New Administrative Capital Stadium, which includes over 6 000 surveillance cameras.** While these features are ostensibly for making life easier and safer, they also grant the authorities an unprecedented ability to monitor public spaces. Scholars argue that this could be used to crack down on citizens wishing to protest or engage in peaceful assembly, thereby threatening basic rights amid a wider clampdown on dissent and freedom of speech[156].

Furthermore, reports from 2019 indicate that the Egyptian government has engaged in cyber espionage by **installing mobile apps on dissidents' phones to extract files, track locations and identify contacts**[157]. The victims of these surveillance activities included journalists, academics, lawyers, opposition politicians and human rights activists, reflecting the government's intent on monitoring and potentially suppressing dissenting voices. The international community, including human rights organisations, has underscored the urgent need for stringent regulation of AI technologies to prevent their misuse for political repression. Allegations suggest that international technological transfers and sales have inadvertently contributed to expanding Egypt's surveillance capabilities, despite foreign governments' criticism of Egypt's practices.

3.6 Algorithmic authoritarianism in Sub-Saharan Africa: The case of Ethiopia and beyond

Ethiopia serves as a novel and interesting case to explore how AI is shaping state-society relations. The country's diverse composition with over 80 ethnic groups provides a critical setting in which to investigate the potential biases and impacts of algorithmic authoritarianism. The use of AI and machine learning in surveillance and policing often exacerbates existing ethnic tensions. Furthermore, the Ethiopian government's control over the country's digital infrastructure, including a monopoly over internet and telecommunications services by **Ethio Telecom,** allows for a unique investigation into state-led algorithmic authoritarianism. Finally, Ethiopia is the second largest country in Africa (population of around 126 million) after Nigeria, albeit having a slightly more diverse ethnolinguistic and sectarian composition. This makes

[154] L. A. Brand, 'Arab uprisings and the changing frontiers of transnational citizenship: Voting from abroad in political transitions', *Political geography*, Vol 41, 2014, pp. 54-63.
[155] D. M. Moss, M. Michaelsen and G. Kennedy, 'Going after the family: Transnational repression and the proxy punishment of Middle Eastern diasporas', *Global Networks*, Vol 22, No 4, 2022, pp. 735-751.
[156] M. Edel and M. Josua, 'How authoritarian rulers seek to legitimize repression: framing mass killings in Egypt and Uzbekistan', *Democratization*, Vol 25, No 5, 2018, pp. 882-900.
[157] M. Josua and M. Edel, 'The Arab uprisings and the return of repression', *Mediterranean Politics*, Vol 26, No 5, 2021, pp. 586-611.

Ethiopia a compelling example through which to explore the **intersection of algorithmic authoritarianism with ethnic and cultural diversity** and serves as a good case study for exploring the effects of AI on state-society relations in a large and diverse society.

Ethiopia has also been **undergoing substantial political transformation alongside rapid digital expansion** as a rapidly developing nation. The government's increasing focus on digitalisation, including the development of digital IDs and a growing online presence, intersects with a political environment that has experienced censorship and surveillance. This juxtaposition provides a fertile ground for studying how emerging technologies might be employed for repressive purposes, especially in contexts of political instability or conflict. While the country has shown a growing interest in AI for positive socio-economic development and technological innovation, there are concerns about how these capabilities might also be used for political repression.

Ethiopia's engagement with AI-**based surveillance and monitoring technologies has raised concerns regarding their use for political repression.** Disclosures by Edward Snowden revealed that the USA provided Ethiopia with surveillance technology and training, which the Ethiopian government may have used to suppress political dissent. The **US National Security Agency set up listening posts in Ethiopia to intercept communications** and provided additional domestic surveillance technology to the Ethiopian army and security agency[158].

The Information **Network Security Agency** in Ethiopia plays a crucial role in facilitating the surveillance of private communications, which has been used to arrest people for lawful opposition activities under the guise of counterterrorism. There are documented instances where transcripts, recordings and phone call metadata were used during violent interrogations and politically motivated trials without judicial warrants[159]. Moreover, Ethiopia has utilised surveillance capabilities obtained from foreign countries, including a Chinese-developed telecommunications system that allows monitoring of every phone call in the country. The government has also used spyware from Italian and German/British firms to target members of the Ethiopian diaspora, reflecting the government's intention to monitor and potentially suppress dissenting voices beyond its borders[160].

The **deployment of surveillance systems** across Sub-Saharan Africa, beyond Ethiopia, has been rising without **sufficient checks, raising fears about repression and the erosion of democratic norms.** Foreign technology, supported by soft loans primarily from China, has increased the accessibility of monitoring products in Africa, including remote-control hacking and eavesdropping capabilities[161]. Such systems enable governments to access files on targeted laptops, log keystrokes and activate webcams as well as microphones. They also allow for the tapping of calls, texts and phone locations, presenting a significant challenge to privacy as well as increasing vulnerability to political surveillance and information suppression.

Recently, researchers have raised concerns about the growing repressive measures used by governments across the continent to suppress digital dissent and civic participation. **The African Digital Rights**

---

[158] D. Grinberg, 'Chilling developments: digital access, surveillance, and the authoritarian dilemma in Ethiopia', *Surveillance & Society*, Vol 15, No 3/4, 2017, pp. 432-438; T. W. Workneh, 'Counter-terrorism in Ethiopia: manufacturing insecurity, monopolizing speech', *Internet Policy Review*, Vol 8, No 1, 2019, pp. 1-22.

[159] T. W. Workneh, 'Digital cleansing? A look into state-sponsored policing of Ethiopian networked communities', *African Journalism Studies*, Vol 36, No 4, 2015, pp. 102-124.

[160] B. O. Dirbaba and P. O'Donnell, 'The double talk of manipulative liberalism in Ethiopia: An example of new strategies of media repression', *African Communication Research*, Vol 5, No 3, 2012, pp. 283-312.

[161] W. H. Gravett, 'Digital neocolonialism: the Chinese surveillance state in Africa', *African Journal of International and Comparative Law*, Vol 30, No 1, 2022, pp. 39-58.

Network (ADRN) conducted a study across ten countries, revealing a two-decade trend of increasing restrictions on online spaces, which were previously considered open fora for free expression and assembly[162]. These tactics include both **overt actions such as internet shutdowns and more covert methods such as online surveillance,** all aimed at limiting freedom of digital expression and assembly. In this report, Juliet Nanfuka from the Collaboration on International ICT Policy for East and Southern Africa, a member of the ADRN, notes a disturbing rise in the use of arbitrary arrests, pervasive surveillance and intimidation to suppress online civic spaces. Financial barriers and regulatory restrictions are spurring self-censorship, eroding fundamental rights to freedom of expression and information. In the face of skyrocketing internet penetration – from a small fraction in 2000 to a quarter of the population in 2019 – the study documents a sobering array of 115 state-controlled internet censorship and control instances across South Africa, Cameroon, Zimbabwe, Uganda, Nigeria, Zambia, Sudan, Kenya, Ethiopia and Egypt. The common thread in these strategies is the deployment of digital surveillance, dissemination of disinformation and calculated internet blackouts, buttressed **by laws that erode digital rights and arrests which silence online speech.** Governments have not shied away from employing AI for targeted monitoring or instituting total internet or mobile phone blackouts, with an uptick in such shutdowns noted in 2020.

4 The role of democracies: Algorithmic bias and technology exports

In discussing algorithmic authoritarianism, it is crucial to capture the global and almost universal nature of a growing problem. Although a focus on authoritarian countries is necessary, autocracies alone are not the sole culprits of algorithmic bias and repression. **Western democracies** whether consciously or inadvertently also contribute to the proliferation of algorithmic authoritarianism capabilities around the world, either by employing these tools domestically or contributing significantly to their export.

It is also worth noting the **economic and market motivations** behind these practices. Investment in advanced AI surveillance technologies can stimulate a country's technology sector, leading to potential economic benefits. By being at the forefront of AI-driven surveillance, countries can position themselves as market leaders, exporting these technologies to other nations, thereby creating a cycle where the tools of repression are both normalised and monetised.

International trade dynamics influencing the **willingness of Western technology companies to aid authoritarian governments** in the development of AI-based repression tools represent a complex interplay of market demands, competitive advantages and geopolitical considerations. At its core, the incentive often revolves around economic gains, but the broader picture can provide more insights into this phenomenon[163].

One of the primary motivations for Western technology companies to engage with authoritarian regimes is the **sheer scale of potential markets** being presented by these countries[164]. Nations with vast populations, such as China or India, present substantial consumer bases and hence significant revenue streams. Engaging with these markets can lead to lucrative contracts that can bolster a company's market visibility

[162] T. Roberts (ed.), Digital Rights in Closing Civic Space: Lessons from Ten African Countries, *Institute of Development Studies*, 2021.
[163] M. Kanetake, 'The EU's export control of cyber surveillance technology: human rights approaches', *Business and Human Rights Journal*, Vol 4 No 1, 2019, pp. 155-162.
[164] M. Ogasawara, 'Mainstreaming colonial experiences in surveillance studies', *Surveillance & Society*, Vol 17, No 5, 2019, pp. 726-729.

and revenues. From an economic standpoint, neglecting these vast markets can be seen as leaving money on the table, especially when competitors might be less scrupulous about such engagements[165].

Moreover, the technology landscape is evolving rapidly, with **innovation being a continual race for the lead.** Western companies are constantly striving to maintain or establish their position as frontrunners in technological advancement[166]. By engaging in partnerships or deals with nations that have substantial resources, these companies can fund their research and development initiatives, ensuring they remain at the cutting edge of innovation. In some cases, authoritarian governments might offer incentives, tax breaks and subsidies to foreign technology companies in exchange for their expertise, thereby making propositions even more enticing[167].

Additionally, by positioning themselves as pioneers in AI-driven surveillance, these companies can **carve out niches in an emerging market segment.** As more nations see the potential benefits (and drawbacks) of such technologies, a demand surge for advanced surveillance solutions is likely. Companies that have established themselves early on can benefit from this demand, having already honed their expertise and solidified their reputation in the field. However, this drive for profit and market dominance can sometimes come at the cost of ethical considerations.

Furthermore, the **monetary benefits reaped by these companies feed into a larger cycle.** As these technologies are normalised, other nations may seek to implement similar systems, seeing them as essential tools for maintaining stability or control[168]. This further drives the demand for advanced surveillance solutions, making the business of repression not just normalised but increasingly profitable. In conclusion, while the lure of vast markets, competitive positioning and the race for innovation all drive Western technology companies towards engagements with authoritarian regimes, the broader implications of these decisions – from ethical concerns to global human rights issues – cannot be overlooked[169]. When left unchecked, the business of AI-driven repression risks creating a world where surveillance becomes the norm, rather than the exception.

For instance, the **USA is a hub for AI innovation,** hosting some of the world's leading technology companies and research institutions. However, it has also been a significant exporter of surveillance technologies such as selling sophisticated facial recognition systems and intercept technologies to countries with questionable human rights records. Domestically, the US has implemented these AI tools in various sectors, including **law enforcement and counterterrorism.** The use of AI in predictive policing and mass data collection initiatives has sparked debates over privacy and the potential for racial profiling and other forms of discrimination.

The EU, despite its strong stance on privacy and human rights, has seen its **Member States exporting surveillance technologies.** These exports often include AI-driven monitoring tools, which in some instances have been acquired by regimes with poor human rights records. The EU's challenge lies in

---

[165] S. M. West, 'Data capitalism: Redefining the logics of surveillance and privacy', *Business & society*, Vol 58, No 1, pp. 20-41. H. A. Ünver and A. S. Ertan, 'Democratization, state capacity and developmental correlates of international artificial intelligence trade', *Democratization*, 2023.

[166] D. H. Flaherty, 'The emergence of surveillance societies in the Western world: Toward the year 2000', *Government Information Quarterly*, 1988.

[167] S. M. West, 'Data capitalism: Redefining the logics of surveillance and privacy, *Business & Society*, Vol 58, No 1, 2019, pp. 20-41; H. A. Ünver and A. S. Ertan, 'Politics of Artificial Intelligence Adoption Unpacking the Regime Type Debate', *Democratic Frontiers: Algorithms and Society*, M. Filimowicz (ed), Routledge (Oxfordshire, England, UK), 2022.

[168] K. Ball, 'All consuming surveillance: surveillance as marketplace icon', *Consumption Markets & Culture*, Vol 20, No 2, 2017, pp. 95-100.

[169] K. Ball, and L. Snider (eds), *The surveillance-industrial complex: A political economy of surveillance,* Routledge (Oxfordshire, England, UK), 2013.

maintaining its ethical standards in AI while being a player in the global technology market. Notably, there have been concerns about European-made surveillance tools being used by authoritarian governments for oppressive purposes, such as tracking and suppressing political dissent[170].

Although **Israel is another significant case,** it will not be a detailed case study within this IDA due to a plethora of studies commissioned by the EP of late, dissecting the involvement of Israeli spyware in EU and international politics[171]. The **NSO Group,** an Israeli technology firm, has gained global attention for developing **Pegasus,** a sophisticated spyware. Pegasus is capable of infiltrating smartphones to access data and conduct surveillance. Although it is not confirmed if Pegasus uses AI, its complexity suggests possible AI elements[172]. The software has been **controversially used to monitor activists, journalists and legal professionals,** often not considered traditional security threats. Investigations by organisations such as Amnesty International[173] and Citizen Lab[174] have demonstrated Pegasus's deployment against civil society members, compromising privacy, press freedom and human rights activities. Pegasus's **capabilities allow deep access to smartphones,** enabling monitoring of conversations, personal data collection and real-time eavesdropping using the device's camera and microphone. This covert surveillance often breaches confidentiality in sensitive cases, impacting press freedom and human rights advocacy. The misuse of Pegasus has sparked debates on surveillance technology regulation and the preservation of civil liberties.

One of the most high-profile cases linked to Pegasus involves Saudi Arabia and the **murder of Jamal Khashoggi,** a journalist who was critical of the Saudi regime. While Khashoggi himself was not directly targeted, his close associates, including a confidante living in Canada, were purportedly monitored before and after his murder in 2018[175]. This case highlights the far-reaching and often hidden impact of such surveillance tools. In **Mexico,** journalists investigating government corruption, such as those reporting on the Iguala mass disappearance, found themselves targeted by Pegasus[176]. **Carmen Aristegui,** a prominent journalist, and her son were among those allegedly spied upon, raising serious concerns about the suppression of press freedom and protection of journalistic sources[177].

The **UAE** also reportedly used Pegasus to target **Ahmed Mansoor,** a human rights activist. Mansoor's phone was hacked through a deceptive text message, leading to his subsequent arrest and highlighting the risks faced by activists in oppressive regimes. **Morocco**'s use of Pegasus against journalist **Omar Radi,** known for his critical views of the government, further exemplifies the software's role in suppressing dissent[178]. In Europe, the use of Pegasus has sparked legal confrontations and political upheaval[179]. **Spain** faced a significant political scandal when allegations surfaced that its government used Pegasus to spy on Catalan separatist politicians. **France** also grappled with Pegasus-related controversies when it was revealed that the spyware might have been used to monitor certain French journalists and possibly even

---

[170] A. Roussi, 'How Europe became the Wild West of spyware', *Politico*, 25 Ocotber 2023.
[171] European Parliament, 'In-depth analysis for the PEGASUS committee, Pegasus and surveillance spyware', Policy Department for Citizens' Rights and Constitutional Affairs Directorate-General for Internal Policies, PE 732.268, 2022.
[172] J. D. Rudie, Z. Katz, S. Kuhbander, and S. Bhunia, 'Technical Analysis of the NSO Group's Pegasus Spyware,' IEE Explore', 2021.
[173] Amnesty International, 'Forensic Methodology Report: How to catch NSO Group's Pegasus', *Amnesty International*, 18 July 2021.
[174] B. Marczak, J. Scott-Railton, and R. Deibert, 'NSO Group Infrastructure Linked to Targeting of Amnesty International and Saudi Dissident', *The Citizen Lab,* 31 July 2018.
[175] D. Priest, 'A UAE agency put Pegasus spyware on phone of Jamal Khashoggi's wife months before his murder, new forensics show', *The Washington Post*, 21 December 2021.
[176] N. Kitroeff and R. Bergman, 'Why Did a Drug Gang Kill 43 Students? Text Messages Hold Clues.', *The New York Times*, 2 September 2023.
[177] P. Ferri, 'Witness in Pegasus case accuses Peña Nieto of ordering spying operation on Carlos Slim', *El Pais,* 5 December2023.
[178] E. Neugeboren, 'Pegasus Spyware Targets Moroccan Journalist', *Voice of America*, 26 June 2020.
[179] D. Boffey, 'EU to launch rare inquiry into Pegasus sypware scandal', *The Guardian*, 10 February 2022.

President Emmanuel Macron. This led to **diplomatic tensions with Israel** and urgent discussions on journalist protection and privacy rights. **Germany**'s revelation that the Federal Criminal Police Office used Pegasus prompted debates within the Bundestag about the balance between surveillance for security purposes and constitutional privacy rights. Similarly, **Hungary, Belgium** and **Poland** have faced scrutiny over the use of Pegasus, with allegations of targeting journalists, members of the European Commission and opposition figures[180]. These incidents have catalysed calls for a comprehensive EU response to regulate such technology[181].

US blacklisting of the NSO Group in November 2021 marked a significant turn in the international stance on private companies supplying spyware capable of transnational repression[182]. This move, restricting NSO Group's access to American technologies, was in response to its alleged role in facilitating human rights abuses across the globe. The blacklisting reflects growing concerns over the use of such technology in oppressive regimes and underscores the need for a global consensus on the ethical use of spyware.

Other Israeli firms such as **Cellebrite, Verint Systems** and **AnyVision,** specialising in digital intelligence and surveillance technologies, have also been scrutinised for their potential role in human rights violations, particularly in Latin America[183]. This has brought to the fore Israel's significant yet complex role in the global cybersecurity and surveillance landscape. The country's vibrant technology sector, known for its advanced surveillance and intelligence tools, faces the challenge of balancing national security interests with ethical considerations and human rights.

Additionally, the Israeli government's **Project Nimbus, involving Google and Amazon, has caused controversy among these multinational companies**[184]. This project, aimed at providing cloud services to Israel, raised ethical concerns among employees due to its potential use in enhancing Israel's digital surveillance in Palestinian territories. Employees feared this could worsen systematic discrimination and displacement. They argued that it contradicted Google's AI principles, which emphasise non-harmful, non-weaponised and norm-compliant AI use. The project's announcement during a period of intense Israel-Palestine conflict, marked by human rights violation accusations, heightened these concerns. This led to significant protests within Google and Amazon, reflecting a growing awareness among tech workers about the societal and ethical impacts of their companies' projects[185].

4.1 US AI-based systems: Concerns over surveillance and privacy

While the USA operates under a democratic system with constitutional rights protecting freedom of speech and expression, there are valid concerns about the use of AI and surveillance tools. Whilst AI is not used to silence dissent in the overt manner of many authoritarian regimes, nevertheless **certain AI-based practices employed by US state governments against migrants and protesters have raised alarms**

[180] J. Rankin, 'EU urged to tighten spyware safeguards in wake of Pegasus revelations', *The Guardian*, 9 May 2023.
[181] European Parliament, 'Committee of Inquiry to investigate the use of Pegasus and equivalent surveillance spyware', webpage, 2023; European Parliament Draft Recommendation to the Council and the Commission following the investigation of alleged contraventions and maladministration in the application of Union law in relation to the use of Pegasus and equivalent surveillance spyware, (2023/2500(RSP)), 22 May 2023.
[182] D. E. Sanger, N. Perlroth, A. Swanson, and R. Bergman, 'U.S. Blacklists Israelie Firm NSO Group Spyware', *The New York Times*, 3 November 2021.
[183] G. Pisanu and V. Arroyo, 'Made Abroad, Deployed at Home', *AccessNow*, 2021.
[184] J. Bhuihan and B. Montgomery, ''A betrayal': Google workers protests Israeli military contract at vigil for ex-intern killed in airstrike', *The Guardian*, 1 December 2023.
[185] Anonymous Google and Amazon workers, 'We are Google and Amazon workers. We condemn Project Nimbus', *The Guardian*, 12 October 2021.

among civil liberties advocates[186]. Accordingly, it is essential to differentiate between targeted attempts to silence political opposition (which are not sanctioned by the state) and broader concerns about mass surveillance as well as potential abuses.

In recent years, the integration of AI into surveillance and policing technologies has sparked a transformative shift in the government's approach to monitoring and managing domestic groups. The motivations for employing such tools often encompass a wide range of objectives, from enhancing national security to ensuring public safety. However, their use has understandably ignited significant debates about privacy, civil liberties and potential misuse. In the USA, the deployment of facial recognition technology has been met with a storm of controversy, stirring debates that touch on the very core of privacy, civil liberties and the risk of perpetuating biases[187]. In a striking instance, the **Immigration and Customs Enforcement (ICE)** searched through driver's licence databases across various states under the cloak of anonymity, pinpointing undocumented immigrants with the cool precision of algorithmic scrutiny[188]. This high-tech tool carries the heavy burden of racial and gender prejudices, as starkly illustrated by the work of cautionary researchers[189]. Their investigations unearthed a troubling propensity for errors, particularly among women and individuals with darker skin tones. The real-world repercussions were unmistakably highlighted in Detroit, where a Black man suffered a wrongful arrest due to a mistaken facial recognition match, casting a long shadow over the technology's reliability in judicial matters[190].

Amidst these challenges, the USA finds itself navigating without guidance from comprehensive federal regulations, allowing facial recognition to proliferate unchecked and unsupervised. The **American Civil Liberties Union (ACLU)**'s exposure of law enforcement agencies adopting these tools without established guidelines or accountability speaks volumes about this regulatory vacuum[191]. Moreover, **the security of vast data repositories underpinning these technologies remains a question,** as evidenced by the 2019 cyberattack that laid bare the sensitive biometric data of thousands of federal agents and officers. This vulnerability not only risks individual privacy but also exposes the potential for data misappropriation to enhance facial recognition capabilities by unknown actors[192]. Concerns escalate when these tools are turned upon the very foundations of democracy, namely public protest and freedom of speech. In the wake of widespread demonstrations sparked by the killing of George Floyd, reports surfaced of law enforcement wielding facial recognition as a sword to identify and track protesters, stirring fears about the right to dissent.

The **use of facial recognition extends beyond the grasp of law enforcement,** penetrating the everyday lives of citizens in schools, apartment buildings and workplaces, often with a murky understanding of consent. For instance, the pushback against a New York school district's intention to use this technology

---

[186] N. T. Lee and C. Chin-Rothmann, 'Police surveillance and facial recognition: Why data privacy is imperative for communities of color', *Brookings Institute,* 12 April 2022.

[187] D. Freeman, D. E. Ho, C. M. Sharkey, and M-F. Cuellar, 'Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies', Report Submitted to the Administrative Conference of the United States, Stanford Law School. February 2022.

[188] C. Edmondson, 'ICE Used Facial Recognition to Mine State Driver's License Databases', *The New York Times*, 7 July 2019.

[189] I. D. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton, 'Saving face: Investigating the ethical concerns of facial recognition auditing', *In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society,* 2020, pp. 145-151; T. J. Benedict, 'The Computer Got It Wrong: Facial Recognition Technology and Establishing Probable Cause to Arrest', *Wash. & Lee L. Rev.,* 79, 2022, pp. 849; A. Tucker, 'The citizen question: making identities visible via facial recognition software at the border', *IEEE Technology and Society Magazine*, Vol 39, No 4, 2022, pp. 52-59.

[190] K. Hill, 'Wrongfully accused by an algorithm', *In Ethics of Data and Analytics, Auerbach Publications*, 2022, pp. 138-142.

[191] M. Nkonde, 'Automated anti-blackness: facial recognition in Brooklyn, New York', *Harvard Journal of African American Public Policy*, Vol 20, 2019, pp. 30-36.

[192] M. Marelli, 'The SolarWinds hack: Lessons for international humanitarian organizations', *International Review of the Red Cross*, Vol 104, No 919, 2022, pp. 1267-1284.

for security purposes underscores the public's unease. Thus, the narrative of facial recognition in the USA is a tapestry woven with threads of innovation, security measures and the safeguarding of fundamental freedoms. As advocates of privacy and civil rights call for tighter reins on these potent technological tools, there is a greater need for empirical studies that explore to what extent American deployment of facial recognition protocols is 'better' or 'less invasive' than more authoritarian states such as China.

Various US police departments have experimented with **predictive policing,** a method that uses algorithms to analyse historical crime data and forecast potential future crime hotspots[193]. Systems such as those provided by **PredPol** deliver police departments with real-time data-driven insights, allowing for pre-emptive deployment of resources. While the intent is to optimise police efforts, concerns arise regarding reinforcing existing biases, given that these algorithms operate based on historical data which might reflect past prejudices. This means that rather than offering an objective view of potential future crime hotspots, predictive policing might simply direct law enforcement back to the same communities that have been over-policed in the past.

A concrete example was seen in **Los Angeles,** where the predictive policing tools were heavily critiqued for leading to **increased patrols in minority neighbourhoods.** These areas were frequently identified by the algorithms as high-risk, but this often worryingly reflected long-standing patterns of enforcement rather than present-day crime statistics[194]. Chicago's attempt at predictive policing, with its 'heat list', faced similar criticisms. The list aimed to pinpoint individuals at risk of being involved in violence or victims thereof. However, this led to increased surveillance of individuals and communities based largely on historical data, as opposed to current behaviour[195]. As a result, certain groups felt unfairly targeted, thereby straining community relations and trust in law enforcement.

The challenge with predictive policing is not necessarily the technology itself, but its implementation without a critical examination of the input data and historical context it represents. If **historical policing bias** is present in the data, predictive policing can replicate and magnify these biases. The challenge facing the USA, therefore, is how to integrate advanced analytics into policing without perpetuating past injustices. For predictive policing to be a constructive part of law enforcement, it must be accompanied by a **critical analysis of data and continual efforts to mitigate any embedded biases.** Only then can the promise of data-driven policing potentially be realised fairly and equitably.

Gait recognition technology in the USA is an emerging frontier in the domain of surveillance. It is a method that identifies individuals based on their distinctive walking patterns. This technology is attractive to surveillance operations because it can be effective at considerable distances where facial recognition systems fail, owing to poor lighting or when faces are obscured or turned away from cameras[196]. This surveillance tool operates by capturing the minutiae of body mechanics. Advanced algorithms process the data points collected from the way a person walks, the stride length, the arm swing, the weight shift and the overall kinetics of the body. By dissecting these elements, technology can create a 'gait signature' unique to each individual.

For instance, in the realm of national security and border protection, gait recognition is being considered for **identifying and tracking individuals deemed as potential threats across expansive areas,** such as

---

[193] S. Egbert, 'Predictive policing and the platformization of police work', *Surveillance & Society*, Vol 17, No 1 and 2, 2019, pp. 83-88.
[194] B. Benbouzid, 'To predict and to manage. Predictive policing in the United States', *Big Data & Society*, Vol 6, No 1, 2019.
[195] R. Richardson, J. M. Schultz, and K. Crawford, 'Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems', and justice, *NYUL Rev. Online*, 2019, p. 94.
[196] R. Caldas, T. Fadel, F. Buarque, and B. Markert, 'Adaptive predictive systems applied to gait analysis: A systematic review', *Gait & posture*, Vol 77, 2022, pp. 75-82.

airports and border crossings, even when facial or other biometric data are unavailable[197]. Despite its potential benefits, gait recognition comes with its share of problematic implications, mainly related to privacy and ethics. Because it is unobtrusive and can be conducted without knowledge or consent, it poses significant concerns about the indiscriminate tracking of individuals. In the USA, there is an ongoing debate regarding the lack of transparency and regulation around the use of such technologies by law enforcement and government agencies[198].

Moreover, similar to other forms of biometric surveillance, the application of gait recognition technology raises **questions about accuracy and the potential for misidentification,** particularly when used in diverse populations. While gait patterns are less likely to change over time compared to faces, the influence of temporary factors such as injury or footwear on the reliability of this technology is **not yet fully understood.** The use of gait recognition also stirs debate on how such surveillance data is stored, protected and potentially shared across agencies or with other entities, raising the spectre of personal privacies being further eroded. Although concrete cases of problems with gait recognition in the USA are not as widely documented as those concerning facial recognition, the mere potential for misuse in widespread public surveillance warrants closer scrutiny[199]. Hence, as technology advances and becomes more integrated into security infrastructures, calls for clear guidelines and robust oversight mechanisms grow louder.

Cell-site simulators, commonly known by the brand name **Stingrays,** are surveillance tools used by various law enforcement agencies across the USA. These devices act as false cell towers, prompting mobile devices within their range to connect with them. Once a connection is established, the simulator can access unique device identification numbers, such as the **International Mobile Subscriber Identity** and **Electronic Serial Number,** and can also triangulate the location of each device with considerable precision[200]. The problematic use of cell-site simulators in the USA arises primarily from concerns over privacy and the legal standards governing their use. For example, these devices can indiscriminately collect data from all mobile devices in the vicinity, not just from a specific individual under investigation. This bulk data collection can ensnare innocent people, gathering sensitive information without their knowledge. Furthermore, law enforcement agencies have often used these devices with minimal oversight or transparency, leading to a lack of accountability[201].

In **Baltimore, Maryland,** the police department acknowledged the use of **Stingrays over 4 300 times since 2007,** according to a 2015 ACLU report[202]. They were used to track stolen phones and locate kidnapping suspects, but concerns were raised about the scope of data collection and the lack of warrants in some cases. The Federal Bureau of Investigation (FBI) has used cell site simulators in counterterrorism efforts. In some of these cases, they operated under a set of internal guidelines that allowed them to use the technology without a warrant in certain national security cases, although the specifics of these operations are often not publicly disclosed due to their sensitive nature[203]. ICE has reportedly used Stingrays to

---

[197] E. J. Harris, I. H. Khoo, and E. Demircan, 'A survey of human gait-based artificial intelligence applications', *Frontiers in Robotics and AI,* Vol 8, 2022.

[198] S. Mouloodi, H. Rahmanpanah, S. Gohari, C. Burvill, K. M. Tse, and H. M. Davies, 'What can artificial intelligence and machine learning tell us? A review of applications to equine biomechanical research', *Journal of the Mechanical Behavior of Biomedical Materials*, Vol 123, 2021.

[199] B. Powell, E. Avidan, and S. Latifi, 'Threat Recognition from Gait Analysis', *IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference*, 2019, pp. 1005-9999.

[200] H. Gee, 'Almost Gone: The Vanishing Fourth Amendment's Allowance of Stingray Surveillance in a Post-Carpenter Age', *S. Cal. Rev. L. & Soc. Just*., Vol 28, 2019, p. 409.

[201] M. Andrejevic, L. Dencik, and E. Treré, 'From pre-emption to slowness: Assessing the contrasting temporalities of data-driven predictive policing', *New Media & Society,* Vol 22, No 9, 2020, pp. 1528-1544.

[202] M. Curtis, 'Aclu Challenges Use Of "Stingray" Surveillance Technology By Baltimore Police', *ACLU Maryland*, 26 November, 2014.

[203] D. Cameron, 'Docs Show FBI Pressures Cops to Keep Phone Surveillance Secrets', *Wired,* 22 June 2023.

locate individuals for immigration enforcement. This use has been contentious, given the broader debate over immigration policy and enforcement in the USA[204]. The Drug Enforcement Administration has used these devices to track suspects in drug investigations[205]. Cases have been documented where data from cell-site simulators have led to arrests and prosecutions for drug-related offences.

The technical capabilities of **cell-site simulators** have evolved, and the potential integration of AI could magnify both their utility and the privacy concerns associated with them. AI can process the vast amounts of data collected by these simulators much faster than human analysts, identifying patterns and connections between users. This could potentially lead to more accurate tracking of suspects' movements and associations. However, it also raises the possibility of more extensive and sophisticated surveillance, exacerbating the challenges of balancing privacy rights with law enforcement needs. **While the technology itself is neutral, how it is being deployed and governed raises critical ethical and legal questions.** The USA continues to grapple with the implications of such surveillance technologies, seeking to find a middle ground that respects the privacy of its citizens while ensuring national security and public safety.

4.1.1 Private companies in US domestic AI-based monitoring systems and American AI Systems Exports

The increasing ubiquity of AI technologies in surveillance and monitoring capacities has been mirrored by the active participation of private companies in developing and providing these solutions. These companies, ranging from Silicon Valley giants to specialised start-ups, offer a suite of AI-driven tools that cater to various law enforcement and government needs. The following list highlights the role and contributions of some key private entities in shaping the US domestic AI-based monitoring landscape.

Palantir Technologies: Founded in 2003, this company has grown to become one of the primary data analytics providers for the US government. Its platforms, notably Palantir Gotham, have been utilised by agencies ranging from the Central Intelligence Agency to the New York Police Department. Gotham's capabilities include data integration from disparate sources and powerful analytics that can identify patterns or links between data points, making it a potent tool for intelligence agencies and law enforcement[206].

Amazon's Rekognition: Amazon Web Services, the company's cloud computing arm, offers Rekognition, a deep learning-based image and video analysis service. It can identify objects, people and even sentiments in images or videos. Law enforcement agencies have explored its utility for tasks such as real-time facial recognition from surveillance camera feeds[207].

Clearview AI: This relatively new player has stirred significant controversy due to its facial recognition tool, which purportedly scrapes billions of images from the internet to build its database. Various law enforcement entities have trialled or employed Clearview's tool for investigative purposes, in recognition of its expansive database[208].

---

[204] US Department of Security, DHS OIG Report: Secret Service and ICE Illegally Used Cell-Site Simulators, Electronic Privacy Information Center, 2023.

[205] S. Hawkins, 'Cell-Site Clone to Track Narcotics Suspect Approved, With Limits',' *Bloomberg Law*, 25 August 2022.

[206] A. Iliadis and A. Acker, 'The seer and the seen: Surveying Palantir's surveillance platform', *The Information Society*, Vol 38, No 5, 2022, pp. 334-363.

[207] S. Mane and G. Shah, 'Facial recognition, expression recognition, and gender identification', *In Data Management, Analytics and Innovation: Proceedings of ICDMAI 2018*, Volume 1, 2019, pp. 275-290.

[208] I. N. Rezende, 'Facial recognition in police hands: Assessing the 'Clearview case' from a European perspective', *New Journal of European Criminal Law*, Vol 11, No 3, 2020, pp. 375-389.

NVIDIA is renowned for its GPUs (Graphics Processing Units), which are crucial for AI processing and deep learning. These GPUs are exported globally, with significant markets in Europe, Asia and other regions. They are used in various industries such as automotive, healthcare and finance as well as by academic and research institutions for AI and machine learning tasks. NVIDIA is at the flashpoint of a US-China trade war over the sale of AI chips[209].

Intel, as a leading semiconductor manufacturer, exports a variety of processors, including those tailored for AI applications, such as the Intel Nervana Neural Network Processors. Their processors are widely used in AI research and development across Europe, Asia and elsewhere, finding applications in data centres and cloud computing services[210].

IBM offers AI solutions through its WatsonX platform and AI-optimised hardware. IBM's AI products and services are used internationally in sectors such as healthcare, finance and retail. Their reach includes countries in Europe, Asia and beyond, assisting with data analysis, AI-driven customer service and decision-making processes[211].

Google (Alphabet Inc.) exports AI software systems through its subsidiary Alphabet Inc. Its Google Cloud AI services and TensorFlow, an open-source machine learning library, are used by businesses and developers worldwide. Google's AI products have a broad international reach, spanning Europe, Asia and other global markets[212].

Microsoft's Azure AI services are offered globally, providing cloud-based AI solutions. These services are utilised by international clients in a variety of sectors, including healthcare, finance and retail, particularly in European, Asian and Middle Eastern markets[213].

It is essential to **note the dual-edged nature of these technologies.** While they undeniably augment the capabilities of law enforcement agencies, offering tools that can enhance public safety, nevertheless concerns surrounding civil liberties, privacy infringements and potential misuse persist. Collaboration between private entities and the government in this arena underscores the need for clear regulatory frameworks and transparency to ensure that the technologies are harnessed responsibly and ethically.

4.1.2 Legislative and judicial check in the USA

Congressional committees have held hearings to delve into the implications of AI use in monitoring and surveillance applications. Legislators have consequently proposed bills aiming to regulate the development and deployment of AI, mandating transparency, fairness and accountability in these systems[214]. Concurrently, the **US judiciary plays a crucial role** in interpreting and applying these laws, providing oversight to ensure that constitutional rights are not compromised. Courts have on occasion been confronted with cases where individuals or groups challenge the use of AI systems, especially when they believe that their rights have been violated[215].

---

[209] Q. Liu, E. Olcott and T. Bradshaw, 'Nvidia develops AI chips for China in latest bid to avoid US restrictions', *Financial Times*, 9 November 2023.
[210] S. M. Khan and A. Mann, 'AI Chips: What They Are and Why They Matter: An AI Chips Reference', *Center for Security and Emerging Technology*, April 2020.
[211] J. Lee, 'IBM unveils new watsonx, AI and data platform', *Reuters*, 9 May 2023.
[212] S. Banjo and D. Ramli, 'Google to Open Beijing AI Center in Latest Expansion in China', 13 December 2017.
[213] A. Kharpal, 'Microsoft's president meets with the Chinese government to discuss AI co-operation', *CBNC News*, 7 December 2023.
[214] L. Brown-Kaiser and S. Wong, 'Sen. Casey rolls out bills to protect workers from AI surveillance and 'robot bosses'', *NBC News*, 20 July 2023.
[215] C. M. Rodgers and J. Obernolte, 'AI's Rise Flags Need for Federal Privacy and Security Protection', *Bloomberg Law*, 6 November 2023.

While far from exhaustive, these cases often grapple with **questions of due process, equal protection, or even First Amendment rights** in the context of algorithmic decisions. However, despite these checks and balances, there are notable limitations. The speed at which AI technology is evolving frequently out–paces the legislative process, making it challenging for laws to stay abreast of the latest developments[216]. Additionally, the technical complexity of AI can sometimes lead to **legislative oversights or gaps,** as not all lawmakers possess a deep understanding of the intricacies involved. From the judicial side, while courts can provide redress in specific cases, they often rely on existing legal frameworks that may not have been designed with the nuances of AI in mind. This can lead to judicial interpretations that, while legally sound, may not fully address the unique challenges posed by AI[217].

The use of AI in monitoring and surveillance applications has led to significant debate and scrutiny within US legal and legislative bodies. **Various congressional committees have held hearings and discussions on this topic,** with notable mentions being the **House Oversight and Reform Committee**'s examination of facial recognition technology *vis-à-vis* its implications on civil rights and liberties[218]. Bills such as the Algorithmic Accountability Act have been introduced, proposing regulatory measures on high-risk AI systems to ensure that they are developed and deployed transparently and fairly[219]. On the judicial front, cases such as the ***American Civil Liberties Union (ACLU)* vs *the FBI*** have highlighted concerns about the agency's use of facial recognition technology and its potential infringement of First Amendment rights[220]. Similarly, the city of Detroit faced lawsuits challenging the wrongful arrest of an individual based on flawed facial recognition matches[221].

These cases and legal challenges underscore the inherent complexities in intertwining algorithmic decisions with civil rights. While these legislative and judicial interventions show promise in addressing AI repression, there are **inherent challenges.** As mentioned earlier, rapid technological advancement can outstrip the pace of legislation and even though some bills are introduced, they face hurdles in passing through both houses and being signed into law. On the judicial side, understanding and navigating the intricacies of AI demands a depth of technical expertise that courts may not always possess. Moreover, the risk of proving bias or harm in an AI system is significant, with algorithms' proprietary nature often shrouding them in secrecy.

4.2 European high-technology exports

The EU has very strong export controls that limit the flow of advanced technologies to authoritarian countries to be used in algorithmic authoritarianism. That said, in the past, there have been cases where European high-technology exports have been repurposed for authoritarian agendas in third countries. Moreover, European companies are still selling advanced surveillance infrastructure to countries with poor human rights records[222]. Nowadays, as surveillance infrastructure largely reinforces AI surveillance capacities, it is impossible to meaningfully separate one from another.

---

[216] B. Wittes, 'A Machine With First Amendment Rights', *Lawfare Institute*, 31 March 2023.
[217] A. Z. Rozenshtein, 'ChatGPT and the First Amendment: Whose Rights Are We Talking About?', *Lawfare Institute*, 4 April 2023.
[218] US government, 'Hearing Wrap Up: Federal Government Use of Artificial Intelligence Poses Promise, Peril', US House Committee on Oversight and Accountability, 15 September 2023.
[219] J. Mökander, P. Juneja, D. S. Watson and L. Floridi, 'The US Algorithmic Accountability Act of 2022 vs. The EU Artificial Intelligence Act: what can they learn from each other?', *Minds and Machines*, Vol 32, No 4, 2022, pp. 751-758.
[220] For a list of ongoing ACLU vs. FBI cases, please refer to: ACLU, 'MediaJustice, et al. v. Federal Bureau of Investigation, et al.', ACLU List of FBI Cases, 2023.
[221] T. McSorley, 'The Case for a Ban on Facial Recognition Surveillance in Canada', *Surveillance & Society*, Vol 19, No 2, 2021, pp. 250-254.
[222] S. Feldstein and B. Kot, 'Why Does the Global Spyware Industry Continue to Thrive? Trends, Explanations, and Responses', Carnegie, 14 March 2023.

Indeed, **European companies have faced scrutiny in the past** for exporting surveillance technologies that have ultimately been used by authoritarian regimes for purposes such as mass surveillance and suppression of dissent. For example, more than a decade ago European companies were implicated in **selling spy tools to authoritarian regimes in countries such as Syria, Egypt and Libya,** following which these technologies were then utilised against journalists, human rights activists and opposition groups to suppress democratic movements, notably since the Arab Spring in 2010[223]. Despite efforts by the EU to implement stricter controls on the export of such technologies, certain Member States, including Sweden, Finland and former members of the United Kingdom, have been influenced by business interests, leading to dilutions in human rights safeguards.

A specific case involves a Swedish company, **MSAB,** known for its involvement in the digital forensics field. MSAB received public EU funding through the flagship technological research programme Horizon Europe (previously Horizon 2020) and was part of the **'Formobile' project**[224]. This project aimed to develop **technology for unlocking mobile devices without user consent and analysing data for criminal investigations.** However, concerns were raised when technology developed under this project was sold to Myanmar's police force, which was under a civilian government at the time, but later became subsumed within military rule, following a coup in 2021. This sale raised questions about the appropriate regulations for exporting surveillance and forensic technology to regions where there is a high risk of abuse.

The **EU's new dual-use regulation,** which came into effect in 2021, aims to tackle emerging technology and prevent sales that could strengthen authoritarian-leaning regimes[225]. It includes regulating 'cyber-surveillance technology' and considering the potential for human rights violations as a key criterion for limiting exports. The Swedish telecommunications giant **TeliaSonera** was implicated about a decade ago for selling high-tech surveillance gear to authoritarian regimes in countries such as Belarus, Uzbekistan, Azerbaijan, Tajikistan, Georgia and Kazakhstan[226]. This technology enabled these governments to spy on journalists, union leaders and members of the political opposition. The equipment provided by TeliaSonera allowed unrestricted monitoring of all communications, including internet traffic, phone calls and text messages. This issue came to light following an investigation by a Swedish news show, revealing the extent to which the technology was used for mass surveillance and suppressing dissent (Source: Electronic Frontier Foundation).

There have been other documented cases where European companies have exported high-technology systems to countries under authoritarian rule, which have then been used for purposes such as mass surveillance and algorithmic authoritarianism. Some notable examples include:

- **Nokia Siemens Networks in Iran[227]:** In 2009, it was reported that Nokia Siemens Networks, a joint venture between the Finnish company Nokia and the German company Siemens, had sold telecommunications equipment to Iran. This technology included monitoring centres that reportedly enabled the Iranian government to perform mass surveillance, intercepting and analysing its citizens'

[223] P. Howell O'Neill, 'French spyware bosses indicted for their role in the torture of dissidents', *MIT Technology Review*, 22 June 2021.
[224] Z. Campbell, C. L. Chandler, 'Tools for Repression in Myanmar Expose Gap Between EU Tech Investment and Regulation', *The Intercept*, 14 June 2021.
[225] Regulation (EU) 2021/821 of the European Parliament and of the Council of 20 May 2021 setting up a Union regime for the control of exports, brokering, technical assistance, transit and transfer of dual-use items (recast), Official Journal of the EU, L 206/1, 11 June 2021; Prior to the 2021 regulation, there was an annual update of the previous dual use Regulation 428/2009 Annexes. See further documents here: Council Regulation (EC) No 428/2009 of 5 May 2009 setting up a Community regime for the control of exports, transfer, brokering and transit of dual-use items (recast), Official Journal of the EU, 5 May 2009.
[226] E. Galperin, 'Swedish Telcom Giant Teliasonera Caught Helping Authoritarian Regimes Spy on Their Citizens', *Electronic Frontier Foundation*, 18 May 2012.
[227] T. Virki, 'Nokia Siemens to ramp down Iran operations', *Reuters*, 13 December 2011.

voice calls, emails and text messages. This capability was allegedly used to suppress and target any opposition during the 2009 Iranian election protests.

- **Hacking Team in various Countries[228]:** The Italian company Hacking Team, known for its surveillance software called 'Remote Control System' or 'Galileo', was reported to have sold its technology to various countries with poor human rights records. This software facilitates the remote monitoring of computers and smartphones. Reports and leaked documents suggested that it was sold to countries such as Saudi Arabia, Sudan and Kazakhstan, raising concerns about its use for internal repression and surveillance.
- **Teliasonera in Central Asia:** The Swedish telecommunications company Teliasonera (now Telia Company) faced criticism for its operations in authoritarian countries in Central Asia, where it was accused of enabling government surveillance. Investigations revealed that the company provided access to its networks to security agencies in countries such as Uzbekistan and Azerbaijan, which could have been used for monitoring and suppressing dissent.
- **Amesys and Nexa in Libya[229]:** The French technology companies Amesys, a subsidiary of Bull SA, and Nexa Technologies provided surveillance equipment to the Libyan government under Muammar Gaddafi. This technology was reportedly used for monitoring Libyan citizens' digital communications, including those of opposition figures, during the Arab Spring uprisings.

These cases highlight the ethical and human rights challenges associated with exporting advanced technological systems to regimes with questionable human rights records. They also underscore the importance of stringent export controls and due diligence to prevent the misuse of such technologies for repressive purposes.

In 2018, **Privacy International** published one of the most comprehensive investigations of the global export market for advanced surveillance systems in the world, revealing significant involvement by British, German, French and Italian companies[230]. These companies had been selling audio-visual surveillance, location monitoring, intrusion (cybersecurity) and forensics services to third countries. All major European companies have been identified as selling interception, intrusion, deep packet inspection and location tracking services to several countries – including autocracies, rendering European involvement in high-technology repression ecosystems as systematic and significant.

In 2020, **Amnesty International** published a report, surveying how European companies, *inter alia* the Dutch **ASML** Holding, a pivotal player in the semiconductor industry, **Ericsson** from Sweden, the Finnish **Nokia** Corporation, Germany's **Siemens** AG, Swedish **Axis** Communications, Czech Republic's **Avast** and Romanian company **Bitdefender** have been involved in the export of high-technology infrastructure to be repurposed for repression and surveillance of dissidents in authoritarian countries[231]. More recently, **Carnegie Endowment's AI Global Surveillance (AIGS)** Index has demonstrated the increasing worldwide interconnectedness of advanced surveillance systems trade[232]. The index and its underlying dataset show a growing market share of EU companies in the global export of advanced surveillance and monitoring infrastructure and systems, including supplies to authoritarian governments. In September 2021, to

[228] A. Greenberg, 'Hacking Team Breach Shows a Global Spying Firm Run Amok', 6 July 2015.
[229] International Federation for Human Rights, 'Surveillance and torture in Egypt and Libya: Amesys and Nexa Technologies executives indicted', Press release, 22 June 2021.
[230] Privacy International, 'The Global Surveillance Industry', webpage, 16 February 2018.
[231] Amnesty International, Out of Control: Failing EU Laws for Digital Surveillance Export, 21 September 2020.
[232] Carnegie Endowment for International Peace, 'AI Global Surveillance Technology', webpage, nd.

address the growing problem strengthened rules on EU export controls entered into force, particularly for dual-use technologies such as advanced computing and spyware[233].

5 Assessing the effectiveness of the current international regulatory framework and governance initiatives on AI

The landscape of international initiatives seeking to create rules and governing agencies for AI is complex and evolving, with multiple actors involved, including the EU, the UN and the CoE together with various other international and regional bodies as well as states and 'blocs' of states (e.g., G7 or the BRICS). Maintaining transparency in the application of AI, especially in sensitive sectors such as security and law enforcement, is essential to avert potential human rights violations. To navigate this terrain effectively, comprehensive and inclusive dialogue is needed, bringing together experts in technology, human rights advocates, policy-makers and public representatives. This collaboration is crucial for thoroughly understanding the implications of AI on fundamental human rights and ensuring that the technology is developed and managed in a responsible, ethical manner which respects the integrity and dignity of individuals.

While there is no universal legal framework governing AI, certain **key international efforts** set the standards and principles to guide its ethical and responsible use. As mentioned afterwards, for example, the UK government recently published its 'Bletchley Declaration' (Section 5. 4. 2), which highlights the global recognition of AI as a transformative force with the potential to enhance human wellbeing, peace and prosperity. It underscores a collective commitment to ensuring AI is developed and used in a manner that is safe, human-centric, trustworthy and responsible. Recognising AI's widespread application across various sectors, the Declaration emphasises the need for its safe development and use for the benefit of all. This approach extends to public services such as health and education, as well as areas such as food security, science and climate change, aligning with the UN Sustainable Development Goals (SDGs). The Declaration also places special emphasis on the safety risks associated with 'frontier' AI, which includes highly capable general-purpose AI models such as foundation models and specific narrow AI capable of causing harm.

The rapid and uncertain rate of AI development, coupled with increased investment in technology, makes understanding and addressing these risks particularly urgent. The inherently international nature of many AI risks is highlighted, with a resolution of working together through international cooperation to ensure that AI is developed in a human-centric, trustworthy and responsible manner[234]. Summarised below are some of the initiatives that are widely discussed in the policy domain and cited most in scientific studies, along with an assessment of their efficacy and limitations.

5.1 The EU

Beyond position statements, the EU is the pioneer in actual, binding norm-setting AI regulations and remains a source of 'best practices' for many other nations that are trying to formulate their own national AI strategies. The EU's strategy for setting international norms involves spearheading and participating in multilateral discussions aimed at establishing a common ethical framework for AI. This is reflected in the EU's active engagement with global institutions such as the UN, where it supports resolutions and contributes to reports that shape the global discourse on AI. However, the consensus-building process is inherently complex, requiring the reconciliation of divergent national policies and priorities. Through bilateral and multilateral diplomacy, the EU engages with third countries to promote adherence to ethical

[233] European Commission, 'Strengthened EU export control rules kick in', Press Release, 9 September 2021.
[234] UK government, 'The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023', Policy Paper, 1 November 2023.

standards in the development and deployment of AI. This diplomatic engagement is manifested in the EU's foreign policy dialogues and international cooperation agreements, which often include provisions for digital governance and human rights.

Firstly, the EU acts as a **convener and leader in international forums** to advance discussions on AI ethics. This includes active participation in the UN, G7, G20 and other multilateral institutions where it can leverage its diplomatic influence to initiate and shape global discourse on the responsible use of AI. By proposing resolutions and guiding principles, the EU contributes to setting global benchmarks that define ethical AI use.

Secondly, the EU **supports the work of specialised agencies such as UNESCO,** which has been working on the ethical aspects of AI. In collaboration with these agencies, the EU helps develop and promote ethical frameworks that align with its own values of democracy, the rule of law and human rights. This includes the backing of initiatives that aim to create a universally recognised body of principles and standards that govern AI development and usage globally.

Thirdly, the EU's norm-setting efforts are also channelled through its **trade and cooperation agreements.** The EU incorporates clauses related to digital rights and AI ethics into these agreements, thereby conditionally tying economic cooperation to the adherence to certain standards of AI governance[235]. This not only promotes ethical AI practices but also encourages the adoption of similar frameworks by trading partners. Furthermore, the EU also advocates for the establishment of a global regulatory framework for AI that includes mechanisms for transparency, accountability and oversight. This framework could potentially set standards for AI audits, ensure data protection and safeguard against algorithmic biases, all of which are pertinent issues in the fight against repressive applications of AI. In addition, the EU can lead by example in continuing to develop and refine its own regulatory environment for AI. The AI Act, for example, is an ambitious attempt to set standards for trustworthy AI within the EU.

Lastly, the EU can also work towards **building coalitions of like-minded countries** that support democratic values and human rights in the digital realm. These coalitions can serve as blocs that endorse and push for the adoption of ethical AI guidelines in international norm-setting bodies, creating a critical mass that can tip the balance in favour of democratic and ethical AI use worldwide.

The EU has positioned itself as a leading proponent of establishing robust regulatory frameworks for AI that prioritise human rights, ethical considerations and security. For many countries seeking to build a regulatory environment and a national AI strategy, the EU provides a benchmark, given that American and Chinese regulatory frameworks are becoming increasingly difficult to adopt due to significant resource and capacity requirements. Furthermore, even US regulations are lagging behind EU standards in terms of human rights[236]. The EU's approach to AI regulation emphasises the need for transparency, accountability and the safeguarding of fundamental rights. This regulatory posture not only influences AI development within EU Member States but also has broader implications for global norms and standards.

[235] EU-Japan Economic Partnership Agreement includes provisions on data protection and the free flow of data, which are closely related to AI ethics, given the importance of data in AI applications. EU-Canada Comprehensive Economic and Trade Agreement (CETA) also includes chapters on digital trade that could provide a framework for future discussions and inclusion of AI-related ethical considerations.

[236] A. Engler, 'The EU and U.S. diverge on AI regulation: A transatlantic comparison and steps to alignment', *The Brookings Institution*, 25 April 2023.

5.1.1 The AI Act

Central to the EU's regulatory strategy is the recently proposed **AI Act,** which aims to create a legal framework for the development, deployment and use of AI across the EU[237]. This act **categorises AI applications based on their risk to citizens' rights and safety,** with a focus on prohibiting high-risk applications that could harm individual or collective rights. By setting **clear standards and obligations for high-risk AI systems,** the EU seeks to ensure that AI technologies are trustworthy and aligned with EU values as well as legal standards. The Act is rooted in a **desire to balance the economic potential of AI with a need to protect fundamental rights and safety,** in essence categorising AI systems based on the level of risk they pose, ranging from *minimal* to *unacceptable* risk. This risk-based approach is crucial, as it allows for a tailored regulatory response, ensuring that high-risk AI systems are subject to stricter controls and requirements, while less risky AI systems face fewer regulatory hurdles. This guards against innovation being stifled by overregulation while safeguarding against misuse of AI in critical areas.

For high-risk AI applications, such as those impacting legal or democratic processes, public health and security, the Act mandates **strict compliance requirements.** These include high standards of data quality, ensuring transparency and traceability of AI systems, as well as implementation of robust human oversight to prevent bad decisions. The aim is to ensure **that AI systems are safe, reliable and respect fundamental rights, including privacy and non-discrimination.** Furthermore, the Act **prohibits certain AI practices** deemed as posing an unacceptable level of risk, such as systems that manipulate human behaviour to circumvent users' free will or government 'social scoring' systems. The EU AI Act also emphasises **transparency, particularly for AI systems that interact with people or are used to detect emotions and categorise individuals.** Users should be aware that they are interacting with an AI system unless disclosure would compromise the system's purpose (e.g., for law enforcement).

The European Commission's proposal for the AI Act marks a **pioneering turning point in the realm of AI regulation**, striving to establish a comprehensive and harmonised binding legal framework across the Union[238]. This ambitious initiative is the first of its kind to attempt a horizontal regulation of AI, addressing the nuanced use of AI systems and the multifarious risks they pose.

The Act proposes a **technology-neutral definition** of AI systems in EU law, aiming to accommodate a wide array of AI methodologies and applications. This inclusivity is pivotal in ensuring that the regulation remains relevant and applicable across the evolving landscape of AI technologies. Central to the Act is its risk-based classification system, distinguishing AI applications as 'unacceptable', 'high-risk', 'limited risk', and 'minimal risk'. This stratification allows for tailored regulatory responses, ensuring that stringent controls are reserved for systems where the potential for harm is greatest.

For AI systems deemed 'high-risk', the Act mandates rigorous compliance requirements **before market entry,** such as: robust risk management and data governance protocols; transparency in operations; and human oversight mechanisms. This emphasis on pre-market evaluation is crucial for mitigating risks to fundamental rights and user safety. Conversely, AI systems presenting 'unacceptable' risks, such as those deploying manipulative subliminal techniques or exploiting vulnerabilities, face outright prohibition under the Act. This bold stance reflects a commitment to uphold ethical AI standards and protect public welfare.

---

[237] European 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, 24 January 2021.

[238] 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, 2021/0106(COD), 26 January 2024.

Soon after the provisional agreement on the regulation in December 2023, several criticisms were presented clustered around five major points:

- First, the Act's broad scope presents considerable challenges in implementation. Ensuring compliance across diverse AI applications, each with its unique technical characteristics and potential risks, requires robust and agile regulatory mechanisms. This complexity is compounded by the rapid pace of AI innovation, which continually tests the adaptability of regulatory frameworks[239].
- Second, given the EU's economic clout, the Act could exert a significant 'Brussels Effect' on global AI markets. Companies outside the EU, especially those keen to access its market, might find it more practical to align their AI products with EU standards, potentially leading to a *de facto* global standardisation in AI development and deployment. According to external experts and analysis, this influence extends to the ethical and safety benchmarks set by the Act, potentially elevating global AI practices[240].
- Third, a critical challenge for the Act lies in striking a balance between fostering AI innovation and ensuring regulatory oversight. Overly stringent regulations risk stifling technological advancement, while lenient measures might fail to address the risks posed by AI adequately. This balancing act is crucial for maintaining the EU's competitiveness in the global AI arena while safeguarding ethical and safety standards.
- Fourth, the Act seeks to harmonise AI regulation within the EU. However, it could contribute to a fragmented global AI regulatory landscape, as different regions may adopt varying standards. This divergence poses challenges for multinational AI developers and users operating across different jurisdictions.
- Finally, the Act's broad language and principles, while ensuring inclusivity, allow for considerable variability in interpretation and application. This variability could lead to inconsistent AI governance practices within the EU, potentially undermining the objective of creating a unified regulatory environment.

In essence, the proposed EU AI Act is a **ground-breaking step towards establishing a cohesive regulatory framework for AI.** However, its effectiveness hinges on the **delicate balance between regulatory rigour and technological innovation,** its adaptability to rapid AI advancements and its impact on global AI markets and practices. From a positive perspective, the Act's approach to AI governance, prioritising safety, fundamental rights and ethical standards, sets a potential blueprint that could influence the development and use of AI technologies globally, shaping the trajectory of AI innovation and its societal integration. The EU is also well-positioned institutionally to tackle some of the Act's early criticisms and can leverage a broad range of expertise to solve some of the initial regulatory problems that may arise.

5.1.2 The Ethics Guidelines for Trustworthy AI

Moreover, the EU has developed the **Ethics Guidelines for Trustworthy AI,** crafted by the High-Level Expert Group on AI[241]. These guidelines set out key requirements for trustworthy AI, including respect for human autonomy, prevention of harm, fairness and explicability. These principles serve as a benchmark

[239] Atlantic Council, 'Experts react: The EU made a deal on AI rules. But can regulators move at the speed of tech?', 11 December 2023.
[240] Siegmann, C. and Anderljung, M. 'The Brussels Effect and Artificial Intelligence: How EU Regulation Will Impact the Global AI Market', *Centre for the Governance of AI*, 16 August 2022.
[241] European Commission, Ethics guidelines for trustworthy AI, 8 April 2019.

for AI development and have the potential to shape international discussions on ethical AI. To complement regulations with up-to-date scientific research, through programmes such as Horizon Europe, the EU is funding research that adheres to its ethical standards, fostering an ecosystem of innovation that contributes to the development of AI and is aligned with its regulatory philosophy.

The guidelines embed ethical considerations into the fabric of AI development and deployment[242]. These guidelines are **globally significant and norm-setting,** ensuring that AI systems are not only technologically proficient but also **adhere to fundamental ethical principles.** Firstly, they have been instrumental in bringing about a shift in the discourse surrounding AI from a predominantly technical to an ethically grounded conversation. By advocating for AI that is lawful, ethical and robust, these guidelines have spurred organisations and developers within the EU and beyond to incorporate ethical considerations at each stage of an AI system's lifecycle. Their influence extends to policy formulation, providing a blueprint for EU Member States to shape their national AI strategies. They have also significantly impacted the corporate sector, guiding companies in developing responsible AI practices and fostering a culture of ethical AI within the business community.

These guidelines recognise the array of challenges and risks associated with AI by addressing the concerns about transparency, bias and fairness, as well as the potential for AI systems to perpetuate existing societal inequalities. By emphasising human agency and oversight, the guidelines seek to ensure that AI systems support human decision-making rather than replace it, thereby mitigating the risks of **dehumanisation.** Another critical focus area is the impact of AI on privacy and data protection, acknowledging the EU's stringent data protection laws, particularly GDPR. The guidelines advocate for privacy-by-design approaches to AI, ensuring that personal data and the privacy rights of individuals are respected.

However, despite their comprehensive and globally norm-setting nature, the EU's Guidelines have been criticised by the policy and corporate domain. As with the OECD (see Section 5.3.2), a primary criticism **lay in their non-binding status,** especially before the approval of the more binding EU AI Act. Without legal enforceability, these guidelines risked being relegated to mere recommendations that lack the teeth to effect real change, especially in the face of economic pressures or technological expediency. Additionally, their principles were found to be **abstract and, at times, challenging to operationalise.** Critics also pointed out the potential for these guidelines to **stifle innovation.** The concern is that stringent ethical requirements might impede technological advancements, especially in a global context where competitors might not be subject to similar **ethical constraints.** Furthermore, the guidelines' focus on human-centric AI has sparked debate. While this focus is lauded for preserving human dignity and rights, there is a discourse on how it might limit the exploration of AI's full potential in areas where human-like decision-making is not paramount. Lastly, they also face the **challenge of keeping pace** with the rapid evolution of AI technologies.

5.1.3 Other EU initiatives

Furthermore, the EU is working to build capacity and share expertise through initiatives such as the **Digital Europe Programme[243],** which aims to advance the digital transformation of Europe's society and economy. It provides funding for high-performance computing, AI, cybersecurity and advanced digital skills, promoting a competitive and digitally skilled EU that can drive the global conversation on AI governance. The EU's comprehensive approach to AI governance – combining regulatory measures, ethical guidelines, research and innovation support, together with international cooperation – reflects its commitment to fostering an environment where AI benefits society while adhering to democratic

[242] European Commission, 'Ethic guidelines for trustworhty AI', 8 April 2019.
[243] European Commission, 'The Digital Europe Programme', webpage, nd.

principles. This proactive stance not only shapes the internal market but also sets a precedent that influences international regulatory efforts and the global governance of AI technologies. The updated 2021 **Coordinated Plan on AI** outlines joint actions for the European Commission and Member States to align policies to enhance AI investment and innovation while ensuring trust and respect for human rights. It sets out concrete actions for collaboration on shaping global norms through bilateral and multilateral partnerships[244].

The EU's **GDPR** (General Data Protection Regulation) although not AI-specific, is a significant part of the AI governance landscape[245]. At its core, this regulation is about protecting personal data, a major source of information for AI, and mandates that any such data used in applications must be processed lawfully, transparently and for legitimate purposes. This requirement ensures that AI technologies respect user privacy and data protection standards. The GDPR requires explicit consent for the processing of personal data. For AI, this means individuals must be informed and consent to their data being used in AI models. This consent must be freely given, specific, informed and unambiguous, promoting transparency on how personal data is to be used.

5.2 The Council of Europe

The **CoE** has been following a parallel trajectory to the EU's broader efforts in regulating AI. It has been particularly proactive in addressing the challenges and opportunities presented by legal hurdles encountered during the use of AI in policy and regulatory decisions. Some of the CoE's initiatives and roles in the realm of AI include:

- The **Ad hoc Committee on AI**[246] was established with a mandate to examine the feasibility of a legal framework for the development, design and application of AI, based on the Council's standards on human rights, democracy and the rule of law. It focuses on ensuring that AI systems are developed and deployed in ways that are not only transparent and predictable but also have adequate safeguards to prevent discrimination and protect fundamental human rights.
- The CoE's **Convention 108 for the Protection of Individuals**[247] about Automatic Processing of Personal Data is the only binding international instrument on data protection. The Council is working to ensure that the principles enshrined in Convention 108 are upheld in the context of AI, particularly regarding personal data protection, as AI systems often rely on vast quantities of data for their training and operation.
- The **European Ethical Charter on the Use of AI in Judicial Systems**[248] is one of the first legal instruments to provide ethical guidelines for the use of AI in judicial systems. It addresses issues of transparency, impartiality, fairness and privacy, recognising the potential impact of AI on due process and the need to safeguard judicial integrity.

---

[244] European Commission, 'Coordinated Plan on Artificial Intelligence', webpage, 2022.
[245] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Official Journal of the European Union, 119, 4 May 2016; cor. Official Journal of the European Union 127, 23 May 2018.
[246] Council of Europe, 'CAHI – Ad hoc Committee on Artifical Intelligence', webpage, 2024.
[247] Council of Europe, 'Data Protection: Convention 108 and Protocols', webpage, 2024.
[248] Council of Europe, European Commission for the Efficiency of Justice (CEPEJ): CEPEJ Euroepan Ethical Charter on the use of artifical intelligence (AI) in judicial systems and their environment, 2024.

- The CoE's Budapest Convention on Cybercrime is the primary international treaty on cybercrime and electronic evidence. The **Cybercrime Convention Committee**[249] examines how AI intersects with cybercrime, including issues around the use of AI in committing offences and as a tool for law enforcement.
- The **Human Rights Guidelines for AI**[250] provide a framework for ensuring AI systems do not infringe on human rights. They are designed to guide the Council's 46 member states in creating legislation or policies that regulate AI applications.

CoE initiatives emphasise a human rights-centric approach to AI governance. They reflect the Council's broader commitment to democratic values and the rule of law, aiming to ensure that AI development in member states and beyond is aligned with these principles. Through these measures, the CoE contributes to setting international norms on AI that seek to respect individual freedoms and the societal implications of AI technology.

By providing guidance and a legal framework, the CoE plays a significant role in influencing how AI is regulated not just within Europe, but also as a benchmark for standards globally, given its position as a leading voice on human rights and legal standards. The effectiveness of these initiatives, though, ultimately depends on their adoption into member states' national law and the political will to enforce them, a challenge common to all international governance efforts.

5.3 Non-binding international initiatives

In the rapidly evolving landscape of AI, international initiatives play a crucial role in shaping international governance frameworks. The following Section delves into the diverse array of non-binding international efforts (UN's facilitation of global dialogue, the OECD's formulation of AI principles and expert forums) aimed at guiding AI development and deployment while upholding ethical standards and human rights.

5.3.1 The United Nations and the UNESCO Guideline Evaluation

The UN does not directly establish AI regulations or a legal framework in the same way that a national government would. Its role is more about facilitating international dialogue, setting broad principles and providing guidance as well as recommendations on ethical and human rights considerations related to AI. However, some key UN initiatives and bodies contribute to shaping the global approach to AI:

- **UN Educational, Scientific and Cultural Organization (UNESCO) Recommendations on the Ethics of AI[251]:** UNESCO has adopted recommendations that provide a global standard-setting instrument on the ethics of AI. These recommendations emphasise respect for human rights and fundamental freedoms, advocating that AI systems must be transparent, explainable and accountable. They also highlight principles such as fairness and non-discrimination, ensuring that other countries or international agencies' regulations adhere to fundamental UN Human Rights.
- **ITU's Focus on Standards and Policies[252]:** The International Telecommunication Union (ITU), a specialised agency of the UN, works on developing international standards, including those related to AI and telecommunications. ITU's AI for Good initiative is a prominent platform for

[249] Council of Europe, 'Cybercrime: Cybercrime Convention Commitee', webpage, 2024.
[250] Council of Europe, 'Commissioner for Human Rights: Artifical intelligence and human rights', webpage, 2024.
[251] UNESCO, Ethics of Articical Intelligence, Global Forum on the Ethics of Artical Intelligence 2024, UNESCO, 2024.
[252] ITU, 'Artifical intelligence for good', webpage, nd.

dialogue and partnership, aiming to identify practical applications of AI to accelerate progress towards the UN SDGs.

- **UN Global Pulse[253]:** This is an innovation initiative by the UN Secretary-General, aiming to harness big data and AI for sustainable development and humanitarian action. Global Pulse works on developing data privacy and ethics standards in AI applications, offering guidelines for UN agencies and their partners.
- **UN Expert Group Meetings and Fora[254]:** The UN hosts various expert group meetings, fora and panels that discuss AI's impact on areas such as human rights, privacy, security, and ethical challenges. Outcomes from these meetings often shape policy recommendations and guide member states in their national AI regulation efforts.

While the UN does not enforce AI regulations, it plays a crucial role in guiding the international community towards responsible and ethical AI development and use. The organisation's efforts are centred around fostering dialogue, sharing best practices and developing guidelines that align with universal values and principles, particularly in the areas of human rights, equity, transparency and sustainability. In December 2023, the UN Secretary-Genera's Advisory Board published **its Interim Report 'Governing AI for Humanity'**[255] advocating for a more cohesive relationship between global norms and the development and implementation of AI. Through a series of 'seven critical functions' (e.g. risk assessment through horizon scanning), the report provides a series of concrete avenues for enhanced accountability and fair representation for all countries in AI-related decision-making processes. The **UN Envoy on Technology,** a role currently held by Mr Amandeep Singh Gill, is an important part of the UN ecosystem on new technologies, as its role progressively encompassed AI and its relationship with both human rights and the SDGs.

To elaborate on a more specific example, **UNESCO's Recommendations on the Ethics of AI,** adopted by member states, represent a landmark effort in establishing a comprehensive ethical framework for AI on a global scale[256]. This initiative seeks to address the ethical implications of AI technologies and ensure that they are developed and deployed in ways that respect human dignity and diversity. The Recommendations' efficacy is rooted in their **global scope** and the authority carried by UNESCO as an influential international body. They provide guidelines on issues such as transparency, accountability, privacy and non-discrimination. Furthermore, one of their key strengths is an **emphasis on the social and cultural dimensions of AI,** advocating for development that is sensitive to the diverse cultural contexts and values across the globe. They identify and articulate various key problems and challenges posed by AI, highlighting its risk of perpetuating biases and discrimination, potentially exacerbating social inequalities. They call for **systems to be transparent and explainable,** ensuring that decisions made by or with the assistance of AI are understandable and subject to scrutiny. Another critical concern addressed is the impact of AI on **privacy and data protection.** Furthermore, they advocate for stringent measures to protect personal data and ensure that the privacy rights of individuals are not infringed by AI technologies. These recommendations also recognise the potential negative impact of AI on labour markets and employment, urging measures not only to mitigate job displacement but also to promote fair and equitable economic outcomes.

[253] UN Global Pulse, website, nd.
[254] UN, Expert Group Meeting on Science, Technology, and Innovation for the SDGs - Meeting of the 10-Member Group of High-level Representatives, the IATT and other experts on high-impact STI4SDSG solutions, in preparation for the STI Forum and the SDG Summit 2023, UN Department of Economic and Social Affiars, 2023.
[255] UN Secretary-General's AI Advisory Body, 'Interim Report: Governing AI for Humanity', December 2023.
[256] UNESCO, 'Recommendation on the Ethics of Artificial Intelligence', 16 May 2023.

Despite their comprehensive nature, the UNESCO Recommendations – as with most other international initiatives – are **non-binding.** This means that the implementation of these recommendations depends on the voluntary action of member states and other stakeholders, which can lead to inconsistencies and disparities in how they are applied globally. There is also a concern regarding the practical implementation of these guidelines. Translating high-level ethical principles into concrete policies and technical standards can be challenging, especially in this field. Dramatic differences in national legislation on how they adopt AI from a legal standpoint also present a problem. Additionally, the effectiveness of the UNESCO Recommendations is contingent on the political will and resources of member states. In regions with limited technological or regulatory capabilities, implementing these guidelines can be particularly challenging.

In regard specifically to cyber-attacks and cyber norms, amidst diplomatic battles at the UN level, the EU and like-minded states proposed in 2020 a **'UN Programme of Action for advancing responsible state behaviour in cyberspace',** a programme which could lead to the establishment of 'a permanent UN forum to consider the use of ICTs by States in the context of international security'[257]. A couple of years later, this led to the adoption of a Resolution to support the establishment of such a Programme after the end of the 2021-2025 UN Open-Ended Working Group[258].

5.3.2 The OECD, non-binding AI principles and the Global Partnership on AI

The OECD plays a critical role in setting international norms for AI through its work on policy guidance, principles for AI ethics and the promotion of international collaboration. Some of the initiatives and contributions from the OECD in this domain include:

- The **OECD's AI Principles**[259] adopted in May 2019 by OECD member countries and several non-members, are a set of recommendations for responsible stewardship of trustworthy AI. These principles were the first international standards agreed upon by governments for the design, development, and deployment of AI. They focus on AI that benefits people and the planet and respect human rights, transparency, and accountability.
- The **AI Policy Observatory**[260] was launched in February 2020, the OECD.AI is an inclusive hub for public policy on AI. It provides data and multi-disciplinary analysis on AI, which helps countries encourage, nurture, and monitor the responsible development of trustworthy AI systems for the human good.
- The **Going Digital Project**[261] provides a holistic approach to understanding and harnessing the benefits and addressing the challenges of digital transformation. The project's recommendations cover policy areas affected by AI, including the labour market, education, innovation, and competition.

[257] Geneva Internet Platform Dig Watch, 'France and partners propose a programme of action for advancing responsible state behaviour in cyberspace', 8 October 2020; Geneva Internet Platform Dig Watch, 'UN OEWG', nd.
[258] See further details about it: UN, 'Programme of action to advance responsible State behaviour in the use of information and communications technologies in the context of international security: draft resolution / Albania, Argentina, Australia, Austria, Belgium, Bulgaria, Chile, Colombia, Croatia, Cyprus, Czechia, Denmark, Dominican Republic, Egypt, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Japan, Latvia, Lithuania, Luxembourg, Malta, Monaco, Netherlands, Norway, Paraguay, Poland, Portugal, Republic of Korea, Republic of Moldova, Romania, Senegal, Slovakia, Slovenia, Spain, Sweden, Switzerland, Tunisia, Türkiye, Ukraine, United Kingdom of Great Britain and Northern Ireland, United Republic of Tanzania and United States of America', 2022.
[259] OECD, 'AI Principles Overview', webpage, nd.
[260] OECD.AI, website, nd.
[261] OECD, 'Going digital project', webpage, nd.

The Global Partnership on Artificial Intelligence (GPAI)[262], while not a UN body *per se*, is an international and multi-stakeholder initiative to support responsible and human-centric development as well as the use of AI, in line with the UN's SDGs. Although the OECD hosts its Secretariat, the UN Secretary-General is still a patron of GPAI, which involves member states and international organisations in collaborative projects.

GPAI's efficacy primarily stems from the need for a comprehensive, multi-disciplinary approach to AI governance[263]. By bringing together experts from academia, industry, civil society and governments, GPAI facilitates a rich exchange of perspectives and experiences. This diversity is crucial in understanding the layered nature of AI and combining engineering and social sciences perspectives of the challenges associated with abuses of AI. One of the key strengths of GPAI lies in its working groups, which focus on areas such as responsible AI, data governance, the future of work and innovation. These groups not only identify best practices but also work towards practical solutions to ethical, technical and governance challenges posed by AI. Moreover, GPAI's emphasis on **leveraging AI for social good,** particularly in alignment with the UN SDGs, marks its commitment to ensuring AI's benefits are globally inclusive. GPAI acknowledges a range of issues associated with AI, from ethical and human rights concerns to technical challenges such as bias and fairness. It recognises the potential for AI to exacerbate social inequalities and the risk of deploying AI systems without adequate transparency and accountability. Importantly, GPAI highlights the need for robust AI governance frameworks, advocating for policies that are both flexible enough to accommodate rapid technological advances and robust enough to ensure ethical as well as societal protections.

However, GPAI's approach has **similar enforceability problems** that can be seen in many other international regulations which fall short of being binding and remain within the parameters of suggestions or proposals. As an entity that operates primarily on consensus and collaboration, GPAI's recommendations and findings do not have the binding force of law, which reduces the direct impact of its work on national policies and corporate practices. Additionally, there are concerns regarding **representation** within GPAI. While it aims for a diverse set of stakeholders, ensuring that all voices, particularly from less developed regions, are adequately represented and heard remains a challenge. This is crucial for developing a truly global approach to AI repression, as the way developing nations adopt AI has the greatest risk of succumbing to repressive and under-regulated practices. Another limitation is the **pace at which it can respond** to the rapidly evolving field of AI. There is a risk that by the time consensus is reached or research completed, the technological landscape may have shifted, rendering some insights less relevant or applicable. Moreover, the effectiveness of GPAI is contingent on the **commitment and active participation** of its member countries and organisations. Differences in priorities and the political will of member states can influence the focus and outcomes of GPAI's initiatives.

The **OECD Principles on AI**[264] **and G20 AI Principles**[265] represent a focused effort to establish a global consensus on the responsible stewardship of AI technologies. These principles, albeit influential, not only embody the inherent complexities and challenges reflective of rapidly developing techniques and capabilities of AI but also the repertoires of repression that emerge as a result of these newer techniques. The OECD and G20 AI Principles are **almost identical in terms of scope,** problem designations and regulatory approach. However, their efficacy predominantly lies in the creation of a guiding framework that transcends national boundaries and aims to universalise some of the key definitions and concepts surrounding what it means to use and deploy an AI-based system. Both principles have been influential in informing the development of national AI ethics guidelines in countries with advanced digital economies and are referenced by businesses and multinational corporations in shaping internal AI governance

[262] GPAI, website, nd.
[263] GPAI, webpage, nd.
[264] OECD, 'AI Principles Overview', webpage, nd.
[265] OECD, 'G20 AI Principles', webpage, nd.

frameworks. In defining the problems associated with AI, they address the quintessential AI conundrum – the balance between **leveraging AI for economic and societal benefits** while safeguarding individual rights and democratic values. Furthermore, they bring into focus the issue of **accountability** in AI, highlighting the need for clear responsibility chains. This aspect is critical in contexts where AI decision-making processes are opaque, making it challenging to attribute accountability in cases of harm or bias.

Their limitations can be clustered around several universal problems associated with the very aim of regulating AI. Their **non-binding nature** is a fundamental constraint: this voluntary approach, while flexible, risks creating a fragmented landscape where the implementation of these principles varies widely, potentially leading to a 'race to the bottom' in ethical standards within competitive global markets. Moreover, the **principles' broad and sometimes ambiguous language,** while necessary for global consensus, may lead to divergent interpretations and applications. This ambiguity poses a challenge for operationalisation, particularly when translating high-level ethical considerations into concrete regulatory or technical actions. Albeit forward-looking, the principles may **struggle to keep pace with rapid advancements** in AI, such as the emergence of more sophisticated machine learning forms, quantum computing influences and novel AI applications could lead to regulatory and ethical oversights. Another critical limitation is the **influence of dominant AI players** – both countries and corporations – in shaping these principles. There exists a risk that the principles may be swayed by the interests of these entities, potentially overlooking the needs and contexts of less represented regions or smaller organisations. Lastly, the **global applicability of the principles is a complex affair** due to cultural and ethical diversity. Ethical AI in one context may not align with the norms or values in another, raising questions about the universality of these principles.

Furthermore, the **G20** has contributed to shaping a global agenda on AI, primarily through recommendations and endorsements of principles and guidelines for ethical AI. Key initiatives and contributions from the G20 in this area include:

- **The Osaka Declaration on Digital Economy[266]:** At the G20 summit in Osaka in 2019, leaders adopted the Osaka Declaration on Digital Economy, which emphasises the importance of international cooperation in fostering trust, security and free-flowing data in the digital economy. It also acknowledges the significance of AI principles for responsible stewardship of trustworthy AI.
- **G20 AI Principles[267]:** The G20 endorsed the OECD's AI Principles, which are a set of guidelines for responsible stewardship of trustworthy AI. These principles include recommendations for AI that respect human rights, democratic values, transparency, robustness, security and accountability. Endorsement by the G20 gave these principles wider international recognition and support.
- **Data Free Flow with Trust[268]:** The G20's emphasis on 'Data Free Flow with Trust' is also relevant in the context of AI. This concept encourages the free movement of data across borders while respecting privacy, data protection and cybersecurity – all crucial elements in the development and deployment of ethical AI.

[266] Japan government, 'G20 Osaka Leaders' Declaration', webpage, 2019.
[267] Japan government, 'Annex. G20 AI Principles 1. The G20 supports the Principles for responsible stewardship of Trustworthy AI in Section 1' and takes note of the Recommendations in Section 2', 2019.
[268] World Economic Forum, 'Data Free Flow with Trust: Overcoming Barriers to Cross-Border Data Flows', White Paper, 16 January 2023.

- **Policy Recommendations on AI:** While not directly issuing regulations, the G20 discussions often result in policy recommendations for member countries regarding the development and use of AI. These recommendations usually emphasise the balance between innovation and ethical considerations, encouraging members to adopt national strategies that align with shared global principles.
- **G20 AI Dialogue[269]:** The G20 regularly facilitates dialogues and discussions among member states on the impact of AI on various sectors, including the economy, labour, healthcare and education. These dialogues help in shaping a collective understanding and approach to addressing the challenges and opportunities presented by AI.

5.3.3 Expert forums

The **Institute of Electrical and Electronics Engineers (IEEE)** is a key player in the AI domain, contributing to the ethical and professional standards within the field. Although a scientific entity, IEEE provides unique opportunities for the dissemination of democratic and ethical norms in AI regulation, due to the large representation of AI scientists from around the world (460 000 members in more than 190 countries), including those from 'authoritarian' regimes'[270].

The IEEE's **Global Initiative on Ethics of Autonomous and Intelligent Systems**[271] exemplifies its commitment to ensuring that ethical considerations are integral to AI development. This initiative promotes education and training that prioritises ethical practices in the design and use of AI systems. A significant document produced by IEEE is the **Ethically Aligned Design**[272], which outlines principles and recommendations developed to inform ethical AI development. This document serves as a guideline for professionals to integrate ethical considerations into their AI projects. The IEEE also has a hand in standards development, particularly through the **IEEE P7000**[273] ('IEEE Standard Model Process for Addressing Ethical Concerns during System Design') series. These standards address various aspects of AI, including privacy, algorithmic bias and transparency, thereby setting the bar for responsible AI development. Education plays a key role in IEEE's activities, with the organisation offering resources and events that focus on the implications of AI in society. Through certification and continual education, IEEE ensures that professionals are updated with current best practices and standards in AI.

The Institute's influence extends into public policy, where it acts as an interface between technology experts and policy-makers, guiding regulatory and legislative developments in AI. Moreover, IEEE collaborates with other standardisation bodies to align AI ethical standards, advocating for a globally consistent approach. This cooperation is part of IEEE's strategy to advocate for AI development that is aligned with human welfare and ethical best practices.

The **IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems** garners input from a diverse array of experts spanning technologists, ethicists, legal scholars and policymakers, which is critical for comprehensively addressing the ethical nuances of AI and autonomous systems[274]. A central achievement of the initiative is the publication of the document **'Ethically Aligned Design' (EAD),** which serves

[269] Saudi government, 'Summary of discussions from the G20 AI dialogue in 2020', 2020.
[270] IEEE, 'IEEE: About at a glance', nd.
[271] IEEE SA, 'The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems', webpage, 2024.
[272] IEEE SA, 'Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomouse and Intelligent Systems', Version 2, 2019, pp. 2-263.
[273] IEEE SA, 'Active Standard: IEEE 7000-2021, IEEE Standard Model Process for Addressing Ethical Concerns during System Design', 15 September 2021.
[274] IEE Standards Association, 'The IEEE Global Initiative', webpage, nd.

as a detailed framework guiding ethical AI development. This document is influential in shaping industry standards and practices, providing a set of guidelines that cover a broad spectrum of ethical concerns in AI.

It examines issues such as algorithmic bias, stressing how biases in data, model design and developer perspectives can lead to skewed AI outcomes with unfair and discriminatory impacts. Furthermore, it confronts the often-opaque nature of AI systems, advocating for **greater transparency** in AI **decision-making processes** and clearer lines of **accountability.** This is particularly crucial in high-stakes domains such as healthcare and criminal justice, where the ramifications of AI decisions are profound.

Simultaneously, the initiative grapples with a balance between autonomy and control in AI systems. It underscores the **ethical implications of high autonomy** in AI, emphasising the necessity for maintaining human oversight to preserve moral responsibility and prevent the dehumanisation of decision-making processes. Additionally, in an era where AI systems voraciously consume data, the initiative places a strong emphasis on **privacy,** advocating for technologies and practices that safeguard individual data rights and autonomy.

Moreover, the IEEE initiative does not shy away from the broader socio-economic implications of AI, such as potential job displacement and inequality. It promotes **a vision of AI that augments rather than replaces human capabilities,** aiming to mitigate the disruptive impacts of AI on the job market and social structures. Despite these ambitious goals, the initiative faces significant limitations and critiques. A challenge exists in bridging the gap between its theoretical ethical frameworks and their practical application. This gap often leaves high-level ethical principles as aspirational goals rather than actionable practices.

Critics argue that without concrete methodologies for implementation, these guidelines may not effectively translate into real-world engineering and business practices. As with other initiatives, another critical issue is the **pace of technological development** in AI. Furthermore, the global applicability of these guidelines is challenged by cultural and societal differences. Critics argue that a **one-size-fits-all approach to AI ethics may overlook the nuances of cultural and societal contexts,** necessitating more region-specific and culturally sensitive guidelines. Moreover, the extent to which the AI industry will adopt these guidelines remains a point of scepticism. The voluntary nature of adherence, especially when ethical standards may conflict with commercial interests, raises questions about the **practical influence of these standards** in the industry.

Lastly, the IEEE's efforts are situated within a broader landscape of numerous entities proposing AI ethical guidelines. Less positively, this multiplicity risks creating a fragmented approach to ethical AI practices, potentially leading to confusion and undermining all existing guidelines' effectiveness.

The **World Economic Forum (WEF)** also plays a significant role in AI, with a series of frameworks, guidelines and initiatives to regulate the space:

- **AI and Machine Learning Framework[275]:** The WEF has developed a framework that outlines key considerations and best practices for deploying AI and machine learning technologies. This framework is intended to guide organisations in implementing AI responsibly, ethically and transparently.
- **Centre for the Fourth Industrial Revolution[276]:** The WEF's Centre for the Fourth Industrial Revolution (C4IR) network works on developing policy frameworks and protocols for emerging technologies, including AI. The C4IR network collaborates with governments,

[275] World Economic Forum, 'A Framework for Developing a National Artifical Intelligence Strategy', 4 October 2019.
[276] World Economic Forum, 'Centre for the Fourth Industrial Revolution', webpage, 2024.

businesses, civil society and experts around the world to co-design and pilot innovative approaches to AI governance.

- **Guidelines on AI Procurement[277]:** In collaboration with various stakeholders, the WEF has developed guidelines for governments on AI procurement. These guidelines aim to ensure that public sector AI deployments are not only ethical, transparent and accountable but also adhere to principles of fairness and inclusivity.
- **Toolkit for Responsible AI[278]:** The WEF has also developed an 'AI C-Suite' toolkit for organisations to ensure responsible deployment of AI. This toolkit includes checklists, guidelines and best practices designed to help organisations assess the ethical implications of their AI applications and make informed decisions about AI deployment.
- **AI Board Toolkit[279]:** Aimed at board members of companies, this toolkit guides how to understand and oversee AI technologies. It includes insights into ethical considerations, risk management and governance structures for AI.
- **Global AI Action Alliance (GAIA)[280]:** Launched by the WEF, GAIA is an initiative that brings together stakeholders from different sectors to accelerate the adoption of inclusive, transparent, and trusted AI globally. GAIA focuses on tangible actions to ensure AI benefits society while mitigating its risks.

5.4 State-led initiatives outside the EU

While the following initiatives are focusing only on three states (the USA, China and India) which hold a significant influence on the AI space, one should not forget to consider other states, such as Brazil and Kenya which are respectively strengthening their national regulation to regulate AI[281].

5.4.1 The USA

The USA plays a pivotal role in shaping the development and governance of AI technologies both domestically and internationally. As home to many of the world's leading AI companies and research institutions, the **USA has a significant influence on setting the direction for AI policies and practices globally.** At the federal level, various initiatives and strategies articulate the US approach to AI governance. These policies emphasise the importance of maintaining American leadership in AI, promoting innovation and public trust in AI technologies, protecting civil liberties and preparing the workforce for an AI future.

The **National Institute of Standards and Technology (NIST)** is tasked with creating standards and guidelines to ensure the reliability, robustness and trustworthiness of AI systems[282]. The National Institute influences both domestic and international standard-setting processes for AI. In addition to standard setting, the USA engages in various bilateral and multilateral discussions and agreements related to AI. Through its diplomatic channels and participation in international fora such as the G7, G20, OECD and UN, the USA works to align international norms and policies on AI. It actively contributes to the development of international principles for AI that reflect American values of openness, reliability and respect for intellectual property and privacy. The USA is also a founding member of the GPAI, through which it collaborates

[277] World Economic forum, 'Guidlines for AI Procurement', 2019.
[278] World Economic Forum, 'Empowering AI Leadership: AI C-Suite Toolkit', 12 January 2022.
[279] World Economic Forum, 'Empowering AI Leadership: AI C-Suite Toolkit', 12 January 2022.
[280] World Economic Forum, 'AI Governance Alliance', webpage, nd.
[281] Akemi Shimoda Uechi, C. and Guimarães Moraes T., 'Brazil's path to responsible AI', 27 July 2023; One Trust Data Guidance, 'Kenya: Bill on Robotics and AI society introduced to National Assembly', 4 December 2023.
[282] National Institute of Standards and Technology, website, nd.

with other countries to advance the responsible use of AI in line with shared democratic values and prevent authoritarian uses of AI.

At the interagency level, the US government coordinates AI policy through entities such as the **National Science and Technology Council's Subcommittee on Machine Learning and AI**[283]. This coordination ensures a consistent and comprehensive US policy stance on AI across different sectors and international engagements. Furthermore, the US leverages its extensive research and development infrastructure to support AI innovation while also ensuring that such advancements are consistent with ethical standards. Federal research agencies such as the **National Science Foundation** fund research into AI ethics, governance and policy[284]. In international trade, the USA includes provisions related to digital trade and AI in its trade agreements, which can influence international norms by setting standards for AI-related intellectual property rights, data flows and privacy[285]. The private sector, with its substantial research and development capabilities, also contributes to the USA's role in AI governance.

The US government often engages with private companies to inform policy and encourage industry-led standards and self-regulation. The collective impact of these efforts positions the USA as a **key player in the international conversation on AI,** promoting a vision that aligns with its **national interests and values** while shaping the global AI landscape in terms of innovation, ethical considerations and governance.

5.4.2 China

China has emerged as a major player in the development and application of AI, through which it has begun to shape its approach to AI governance, both within its borders and on the international stage. Although regarded as an authoritarian user of AI, the Chinese government has nonetheless expended considerable effort in domestic and international norm-setting practices and for a to craft its own way for the responsible use of AI. Domestically, China has articulated its ambition to become a global leader in AI through a series of state-led plans and directives. The **'New Generation Artificial Intelligence Development Plan',** launched in 2017, sets out a roadmap for China to become the world leader in AI by 2030, with goals spanning research and development, policy support, talent cultivation and ethical norms[286]. In terms of regulation, the Chinese government has gradually introduced a regulatory framework for AI that encompasses both the promotion of AI and the management of potential risks. The government has issued guidelines and principles for the responsible development of AI, focusing on aligning AI with social values, ensuring security and controllability, promoting transparent and fair principles, as well as encouraging collaboration between government, industry, research institutions and users. The **State Council of China** has also issued various AI governance guidelines that emphasise the importance of ethical standards, intellectual property rights and user privacy[287]. These guidelines aim to foster an environment where AI can flourish while ensuring that it remains within the bounds of Chinese law and policy goals.

Internationally, China has been actively involved in multilateral fora related to AI. It has engaged with organisations such as the UN and the World Trade Organization, seeking to influence global discourse on

---

[283] Office of the President of the USA, 'Charter of the Machine Learning and Artificial Intelligence, Committee on Technology, National and Science Technology Council', 2016.
[284] US National Science Foundation, 'Advancing Ethical Artificial Intelligence Through the Power of Convergent Research', nd.
[285] E. Jones, 'Digital disruption: artificial intelligence and international trade policy', *Oxford Review of Economic Policy*, Vol 39, No 1, Spring 2023, pp. 70–84.
[286] G. Webster and L. Laskai, 'Full Translation: Chinese Expert Group Offers 'Governance Principles' for 'Responsible AI'', DigiChina, Stanford University, 17 June 2019.
[287] S. Larsson, 'On the governance of artificial intelligence through ethics guidelines', *Asian Journal of Law and Society*, Vol 7, No 3, 2020, pp. 437-451.

AI and contribute to the development of international norms. China's participation in these fora is indicative of its intention to play a significant role in shaping the governance of AI technologies worldwide. Additionally, China is a member of the **GPAI** to guide the responsible development and use of AI, consistent with human rights, inclusion, diversity, innovation and economic growth. As part of this group, China contributes to the discussions and knowledge-sharing on AI policy and practices.

Through its bilateral ties, China also **exports its AI technologies and governance approaches** to other countries, which may influence how these nations develop their own AI policies. This export includes not just the technology itself, but also the norms, standards and regulatory principles that accompany its deployment[288]. Despite its active role, China's approach to AI regulation is often viewed as being at odds with Western perspectives, particularly regarding the central role of the state in governance, given not only its handling of personal data but also its emphasis on surveillance and security. China's AI initiatives tend to reflect **broader strategic priorities,** including national security and economic development, which have significant implications for how AI is regulated and used globally. In summary, China's role in regulating and setting international norms on AI is characterised by its **ambitious national AI development plans, emerging regulatory landscape, active participation in international bodies** as well as the **global export of its AI technologies** and **governance principles.** The Chinese approach intertwines AI development with state interests, which has a profound impact on its AI governance practices and the international norms that it advocates.

Certain guidelines and principles have been released by various arms of the government and industry associations that provide insight into China's stance on the ethical and responsible use of AI:

- In June 2019, the **Beijing AI Principles** were published by the Beijing Academy of AI, which outlined a set of principles for the research, development, use, governance and long-term planning of AI, emphasising the need for AI to be beneficial to humanity and the environment[289].
- In 2021, the **Cyberspace Administration of China** released draft regulations to curb the misuse of algorithmic recommendation technologies. The proposed rules aimed to stop practices that may manipulate users' behaviour or spread misinformation[290].
- The **Personal Information Protection Law,** effective from November 2021, while not exclusively focused on AI, is critical to AI ethics as it governs the handling of personal data, which is central to AI systems. It stipulates requirements for data processing transparency, user consent, and data security that AI developers and operators must comply with[291].
- The **Data Security Law,** also effective in 2021, provides a regulatory framework for data security and management, affecting how AI can use and process data[292].
- The **Chinese Ministry of Science and Technology** commissioned an AI governance committee to draft ethical guidelines for AI, and in 2019, the committee released the 'Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence', which advocates for AI to be controllable, transparent, lawful, and ethical[293].

---

[288] H. Roberts, J. Cowls, J. Morley, M. Taddeo, V. Wang and L. Floridi, 'The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation', *AI & society*, Vol 36, 2020, pp. 59-77.

[289] International Research Center for AI Ethics and Governance, 'Beijing AI Principles', webpage, nd.

[290] L. Hurcombe, H. Yong Neo and D. Wong, 'China's cyberspace regulator releases draft measures for managing generative AI services', DLA Piper, Loxology, 18 April 2023.

[291] Deloitte, 'The China Personal Information Protection Law (PIPL)', May 2021.

[292] US government, 'China's data security law', International Trade Administration, 17 August 2021.

[293] G. Webster and L. Laskai, 'Full Translation: Chinese Expert Group Offers 'Governance Principles' for 'Responsible AI'', DigiChina, Stanford University, 17 June 2019.

The EU and various major countries including the UK, the USA, China and Australia have acknowledged the potentially catastrophic risks posed by rapidly evolving AI technology. At the **AI safety summit in November 2023** hosted by the British government, 28 countries, including China, signed the Bletchley Declaration, as part of the first international announcement committing to collaborative efforts on AI safety research. This agreement comes despite apparent competition between the USA and the UK over who should lead the development of new AI regulations. The summit marked a diplomatic success for the UK, particularly for Prime Minister Sunak, who initiated the event amid concerns over the rapid and unregulated advancement of AI models. The summit featured a rare display of global unity, with US Commerce Secretary Gina Raimondo and Chinese Vice-Minister of Science and Technology Wu Zhaohui sharing the stage. China's participation in the declaration was significant, with Wu emphasising principles of mutual respect, equality and mutual benefits in AI development as well as usage[294].

5.4.3 India

India has been working towards establishing a framework for AI that aligns with its digital economy goals and is addressing the ethical dimensions of this technology. While India's role in setting international norms on AI is still emerging, the country has been involved in various initiatives and policy formulations to regulate AI both domestically and internationally. The main governmental public policy think-tank **National Institution for Transforming India Policy Commission** 2018 discussion paper, titled '*#AiforAll: Harnessing the AI for Inclusive Growth*'[295] sets out ethical, legal and societal implications for AI and proposes establishing an Ethics Committee. While this is a strategy document and not law, it nevertheless provides a framework for thinking about how AI should be governed. The **proposed Draft Personal Data Protection Bill** modelled on the EU GDPR, includes principles that would be essential for ethical AI, such as consent, data minimisation and individual rights concerning automated decisions. Although this bill specifically targets data protection, its principles are crucial for the responsible development and application of AI technologies that process personal data.

An **AI task force constituted by the Ministry of Commerce and Industry** recommended the formation of an inter-ministerial **National AI Mission.** It was suggested that this body should play a role in developing and enforcing ethical, legal and regulatory frameworks for AI. Although not AI-specific, the IT Act 2000 and its associated rules provide a legal framework that impacts how AI systems should process data, ensuring certain levels of protection for digital information and transactions[296].

Internationally, **India is an active participant in global fora** where AI norms are debated and shaped, such as the G20 Digital Economy Task Force. The country has also engaged with the World Economic Forum's Centre for the Fourth Industrial Revolution to co-design new policies related to AI and data utilisation. Furthermore, India supports the multi-stakeholder approach in the global governance of AI and has participated in the development of norms through platforms such as the GPAI. Through such international collaborations, India contributes to the global dialogue on responsible AI, sharing its perspectives and expertise.

At a **bilateral level,** India has engaged with China, the USA and the EU, as well as leading AI innovators such as Japan and South Korea, to promote an open and diverse digital economy, including discussions on AI. These engagements help not only in setting a shared understanding of ethical AI but also in fostering innovation and ensuring economic benefits.

[294] K. Stacey and D. Milmo, 'UK, US, EU and China sign declaration of AI's 'catastrophic' danger', *The Guardian*, 1 November 2023.
[295] India government, National Strategy AI For All, June 2018.
[296] Indiacode, 'The Information Technology Act, 2000', 2000.

6 Key recommendations

As AI technologies become increasingly integrated into various aspects of governance and daily life globally, there is a corresponding rise in the potential for their misuse. Authoritarian and illiberal governments may harness these tools for repressive purposes, such as surveillance, censorship and undermining democratic processes. Such actions pose significant risks to human rights and the foundational principles of open societies.

The EU is uniquely situated as a norm-setter in human rights, democracy and the rule of law. As such, it can lead the global discourse and action on ethical AI use. In response to the challenges being presented, a set of comprehensive and strategic recommendations is proposed for the EU and the EP to consider. These recommendations aim to curb the potential for AI abuse and promote the ethical development and deployment of AI technologies. The proposed measures encompass legal, technological, diplomatic and economic aspects, reflecting the multifaceted nature of AI and its implications for society. By taking a proactive stance, the EU can set a global standard for AI ethics and human rights, while also protecting individuals from digital authoritarianism both within and outside its borders. This is crucial for countries that are currently adapting AI in political and legal decision-making and are seeking a 'third option', other than American and Chinese regulatory frameworks that are undesirable or otherwise difficult to copy (see Section 3.2 and 4.1).

In the context of international policy and regulation, the EU has devised a multifaceted approach to mitigate risks associated with the misuse of AI by authoritarian regimes. This approach is informed by a range of EU institutional reports, resolutions from the EP, as well as academic and think tank literature.

6.1 Recommendations for the EU

Sanctions constitute a more direct and punitive measure, targeting entities involved in human rights abuses facilitated by AI technology. The **EU Global Human Rights Sanctions Regime,** which enables the bloc to target individuals and entities responsible for or associated with serious human rights violations, is an example of the EU's commitment to countering the global challenge of AI misuse. On the preventative side, the EU employs export controls to limit the dissemination of AI technologies that could be repurposed for repressive aims. Regulations such as the **EU Dual-Use Regulation** are periodically updated to encompass emerging technologies, reflecting an awareness of the dynamic nature of technological proliferation.

Targeted sanctions: The EU could implement targeted sanctions against specific individuals, companies, or government entities that are found to be responsible for developing, selling, or using AI technologies for repressive purposes. This can include travel bans, asset freezes and other financial sanctions.

Sectoral sanctions: These sanctions would target entire sectors that contribute to the authoritarian use of AI, such as advanced computing, facial recognition technology, or surveillance equipment. By imposing restrictions on these sectors, the EU could disrupt the supply chains and limit the availability of tools used for repression.

- **International collaboration:** The effectiveness of sanctions could be greatly enhanced through collaboration with international partners. The EU could further work with allies to ensure that sanctions are implemented multilaterally, which would increase their impact and minimise the risk of circumvention.
- **Monitoring and enforcement:** To ensure that sanctions are effective, the EU would further need robust mechanisms for monitoring compliance and enforcement. This could involve, *inter alia*, intelligence-sharing agreements, satellite monitoring and AI-driven analysis of financial transactions.

- **The EU's use of sanctions:** to mitigate the authoritarian use of AI, sanctions would serve as a deterrent to potential human rights abuses while also promoting ethical standards in AI development and deployment globally. However, sanctions must be carefully calibrated to avoid unintended consequences, such as exacerbating the plight of the local population or hindering positive AI innovation. It is also crucial for sanctions to be part of a broader strategy that includes diplomatic engagement and international cooperation to promote the adoption of ethical AI practices globally.
- **Conditional access to EU markets:** Access to the EU's lucrative markets could be made conditional upon compliance with international human rights norms, including the ethical use of AI. This would incentivise foreign companies and governments to adhere to these norms to retain market access. The recently adopted EU AI Act mostly covers this issue and introduces a layered set of conditions on the ethical use of AI for entry into the European market.
- **Legal frameworks:** To enable these sanctions, the EU would need to ensure that its legal frameworks are equipped to address the nuances of AI technologies and how they can be misused. This could involve updating existing sanctions regimes to improve the capture of AI-specific considerations.

In terms of capacity building, the EU provides technical assistance to civil society organisations and human rights defenders. Initiatives such as the **Human Rights and Democracy Thematic Programme under Global Europe NDICI** offer financial support and training to enhance their capabilities in monitoring and countering repressive uses of AI. The challenge here is to ensure that such assistance is effective and reaches the right actors without exacerbating their vulnerabilities. While the EU's approach is comprehensive, the efficacy of these policy instruments is contingent upon various factors, including international collaboration and the willingness of third countries to align with the proposed norms:

- **Technical training:** The EU could offer technical training to civil society organisations, journalists and human rights defenders in authoritarian countries to help them understand AI technologies and the ways these can be misused. This knowledge enables these actors to better advocate for responsible AI use and recognise signs of digital repression.
- **Legal expertise:** By providing legal expertise and support, the EU could help countries draft and enforce laws and regulations that govern the ethical use of AI. This can include assistance in creating data protection laws, privacy standards and oversight mechanisms for surveillance technologies.
- **Research collaboration:** Supporting joint research initiatives between European and international universities and think tanks on the ethical use of AI could promote a deeper understanding of how AI can be used responsibly. These collaborations could also develop best practices and guidelines that can be adopted by governments worldwide.
- **Public awareness campaigns:** Funding and organising public awareness campaigns about the potential misuse of AI could empower citizens in authoritarian regimes to demand more transparency and accountability from their governments regarding the deployment of AI systems.
- **Digital infrastructure:** Investing in the digital infrastructure of countries at risk could reduce their dependence on authoritarian states that may offer technology with strings attached. Infrastructure that promotes open and secure internet access is crucial for democratic engagement.

- **Policy advice and consultation:** The EU could advise governments that wish to implement AI technology in ways that are ethical and respect human rights. This could involve the sharing of best practices, policy frameworks and ethical guidelines for AI.
- **Developing open-source tools:** Supporting the development and dissemination of open-source AI technologies could provide alternatives to proprietary tools that may be used for surveillance or censorship. This also includes providing tools that could detect and counteract state-sponsored AI misuse.
- **Promoting AI literacy:** Educational initiatives could be funded to increase AI literacy across various sectors of society, ensuring a broad understanding of AI's benefits and risks. An informed society is better equipped to challenge and debate the introduction and use of surveillance technologies.
- **International partnerships:** Strengthening international partnerships for AI governance could help to promote standards and practices that safeguard human rights. The EU could strengthen its participation or leadership in multilateral fora aimed at creating a common understanding of ethical AI use.
- **Incubators and innovation hubs:** Supporting the creation of incubators and innovation hubs in developing countries could foster the development of local, ethical AI solutions tailored to the specific needs and challenges of these environments.

The EU's capacity-building efforts would be designed to ensure that the global evolution of AI remains aligned with the values of democracy, human rights and the rule of law. By fostering a broader, inclusive and well-informed stakeholder base, the EU could contribute to the resilience of societies against the misapplication of AI technologies by authoritarian regimes.

Through these recommendations, the EU would not only be safeguarding fundamental rights but would also be fostering a technological ecosystem where innovation thrives alongside ethical considerations and respect for individual liberties. Recommendations are intended to be dynamic and adaptable, enabling the EU to respond effectively to the rapidly evolving technological landscape. The overarching goal is to ensure that AI serves the public good, contributes to human well-being and upholds the dignity and rights of all individuals, regardless of the geopolitical context. It is with this vision in mind that the following detailed recommendations are outlined.

Build on the AI Act: Following intensive negotiations, the Council presidency and the EP have provisionally agreed on the AI Act on 9 December 2023, which is a significant step in establishing harmonised rules for AI within the EU. This draft regulation is designed to ensure that AI systems used or marketed in the EU are safe, respect fundamental rights and align with EU values. It also seeks to foster AI investment and innovation in Europe. There was recently a debate between French President Emmanuel Macron and the EU Commissioner for Competition Margrethe Vestager whereby President Macron criticised the AI Act for being 'too strict' and having the potential to disrupt AI innovation across Europe[297]. Vestager in turn, defended the Act by arguing that fears about AI regulation stifling AI innovation in the EU are unfounded. She further asserted that a more streamlined regulation with clearer parameters would support a more rapid AI development, once the general regulation parameters are set in place and companies have a clear sense of what they can and cannot do[298].

To address the concerns raised and to balance the tension between regulation and innovation, the EU AI Act offers the following strategic enhancements:

[297] Le Monde, 'Macron argues against 'punitive' AI regulation', 17 November 2023.
[298] J. Espinoza, 'EU competition chief defends Artificial Intelligence Act after Macron's attack', *Financial Times*, 29 December 2023.

- **Flexible and adaptive regulatory framework:** The EU ensured that the AI Act remains flexible and adaptive to the rapidly evolving nature of AI technologies. This involves establishing mechanisms for regular review and updates of the Act, allowing it to stay current with technological advancements and market trends. Such adaptability ensures that the regulatory framework does not become outdated, thereby supporting ongoing innovation.
- **Clearer guidelines and streamlined compliance:** By providing clearer, more detailed guidelines on compliance, the EU reduced uncertainty for AI developers and businesses. Simplifying the compliance process, especially for SMEs and start-ups, lowers barriers to entry and stimulates innovation. Clear guidelines also help companies understand the boundaries within which they can innovate, speeding up the development process.
- **Promoting public-private partnerships:** Encouraging collaborations between public institutions and private sector entities in AI research and development is a vital step. These partnerships leverage the strengths of both sectors – the regulatory and ethical oversight of public bodies and the agility and innovation of private companies. Such collaborations lead to the development of AI applications that are both innovative and aligned with regulatory standards.
- **Incentives for ethical AI development:** The EU AI Act introduced incentives for companies that prioritise ethical AI development. This includes tax breaks, grants, or other financial incentives for projects that align with the EU's ethical standards and fundamental rights framework. Incentivising ethical AI development could encourage companies to innovate within the regulatory framework, rather than pushing the boundaries of compliance.
- **Fostering AI research and talent development:** Investment in AI research and the nurturing of AI talent within the EU could help in advancing innovation. This could involve funding AI research initiatives, supporting universities and research institutions as well as creating programmes to attract and retain AI talent. A strong research base and a skilled workforce are crucial for sustainable AI innovation.

Promote global regulatory convergence: The EU's GDPR has set a global benchmark for data protection and privacy, offering a template for international regulatory convergence on the governance of AI. To mitigate still further the risk of AI being utilised as an instrument of repression, the EU should aim to promote the harmonisation of AI regulations on a global stage. The following points expand on this policy option:

- The EU should lead in **fostering global standards that enable interoperability of AI systems across borders.** Interoperability ensures that AI systems respect human rights standards universally, not just within the EU. This includes the alignment of technical standards, data protection norms, and ethical guidelines. The EU could leverage the international impact of the **GDPR** as a starting point for advocating similar principles in the governance of AI globally. This involves promoting principles of transparency, consent, data minimisation and individual rights in the context of AI.
- **Active diplomatic efforts** should be directed toward engaging with international partners and multilateral organisations to promote regulatory convergence. The EU could use its influence in international bodies like the UN, the G7 and the G20 to advocate for a consensus on human-rights-respecting AI regulations. Working with international standards bodies, such as the International Organization for Standardization to develop and promote AI-specific standards that are globally recognised and integrated into the regulatory frameworks of member states.

- The EU could **assist countries that may lack the resources to develop their regulatory frameworks** by providing expertise and support, thus ensuring these countries are not left behind in the establishment of AI norms that respect human rights. The EU should seek to include AI governance clauses in trade agreements and international treaties, binding partners to maintain the agreed-upon human rights standards in their AI practices.
- **Develop international mechanisms that facilitate the cross-border enforcement of regulations,** ensuring that entities that operate internationally can be held accountable to these standards regardless of where they are based. Establish a body or enhance the mandate of an existing institution to monitor and report on global trends in AI governance. This could serve to identify areas where convergence is occurring and areas where divergence needs to be addressed through diplomatic and economic channels.

Enhance export controls: To effectively curb the misuse of AI technology for repressive purposes, the EU strengthened its export control regime through the Economic Security Package that was released in January 2024[299]. The following points elaborate on enhancing export controls as a policy measure:

- **Develop a comprehensive list that identifies and categorises sensitive AI technologies and components** that could be repurposed for surveillance, censorship, or repression. This list must be continually updated to keep pace with technological advancements. Establish robust risk assessment protocols to evaluate the potential human rights impacts of exporting AI technologies. These protocols should consider the political climate, rule of law and human rights records of the recipient countries.
- **Amend existing EU export regulations to specifically address AI technologies.** This could involve revisions to the EU Dual-Use Regulation to incorporate a wider array of AI systems that could be misused in the context of human rights abuses. Implement stringent licensing requirements for companies seeking to export AI technology. The EU should be willing to deny export licences when there is a substantial risk that the technology will be used for repression or to commit human rights abuses.
- **Mandate transparency in the export licensing process,** requiring companies to report on the end-use and end-users of AI exports. Considering the **Wassenaar Agreement,** implementing a mandate for increased transparency in the AI export licensing process is crucial. This would involve requiring companies to provide detailed reports on the end-use and end-users of AI exports. Such transparency is essential for effective monitoring and accountability, ensuring that AI technologies are used responsibly and in line with the EU's ethical standards. The EU can leverage the Wassenaar Arrangement framework to collaborate with international partners in establishing specific controls for AI technologies. The Wassenaar Arrangement, which already encompasses a wide range of dual-use goods and technologies, can be a platform for the EU to advocate for and develop a specialised export control regime for AI technologies.
- **Develop mechanisms for the verification of the end-use of exported AI systems,** ensuring they are not being used for purposes other than those stated at the time of export.
- **Sanctions for non-compliance:** Establish clear and stringent penalties for entities that violate export controls, including fines and restrictions on future exports, to ensure compliance and deter illicit trade in sensitive AI technologies.

[299] European Commission, Communication from the Commission to the European Parliament and the Council: Advancing European economic security: an introduction to five new initiatives, COM(2024) 22 final, 24 January 2024.

To mitigate any authoritarian use of AI technologies, the EU can operationalise **its technical assistance to vulnerable states** through a multifaceted approach that combines legislative advisory services, infrastructure support and capacity-building programmes.

- Initially, the EU could establish a **dedicated task force within the EEAS** that specialises in AI governance. This task force would be responsible for the coordination of technical assistance activities, ensuring they are aligned with the EU's foreign policy objectives and the specific needs of partner states. It would work closely with local authorities to assess existing legal and institutional frameworks and identify gaps where AI could be misused for repressive ends.
- **Legislative support from the EU** would probably involve advising on drafting comprehensive data protection laws that mirror key elements of the EU's GDPR, such as consent, data minimisation and individuals' rights concerning their data. The EU could also share its best practices for impact assessments of AI deployments, particularly in sensitive areas such as public surveillance and profiling.
- On the infrastructure front, **the EU could facilitate the establishment of secure data processing facilities** that are resilient to intrusion and unauthorised access. This includes the use of encryption technologies, secure hardware provisions for data storage and advanced network security protocols to safeguard against malicious AI applications that compromise privacy and personal freedoms. Capacity building would extend to organising workshops and training modules for local regulators, law enforcement and civil servants, focusing on the ethical use of AI, the risks associated with machine learning algorithms and the importance of maintaining human oversight in automated decision-making processes. This would be supplemented with knowledge-sharing initiatives, bringing in EU experts to collaborate on research and development projects that reinforce ethical AI innovation within these states.
- Furthermore, **the EU could support the establishment of centres of excellence in partner countries,** which would serve as hubs for the development of ethical AI practices and education. These centres would facilitate the exchange of knowledge between EU and local experts, provide training for stakeholders and advocate for AI systems designed with a human-centric approach that prioritises individual rights and freedoms. This 'technical assistance package' would emphasise building enduring institutional capacities rather than offering one-off training sessions or consultations. It would be aimed at creating sustainable expertise and infrastructures within partner countries, enabling them to uphold the principles of democratic governance independently in the face of emerging AI challenges.

Implement AI auditing and certification: The establishment of an EU-wide AI auditing and certification mechanism constitutes a critical step toward ensuring that both domestic and international deployments of AI technologies adhere to stringent ethical standards and respect human rights. This mechanism would serve as a verification system to assess and certify that AI products and services are developed and used in compliance with established norms.

- Such a mechanism would entail the **development of a comprehensive set of criteria and standards for what constitutes ethical AI use.** These criteria would be grounded in the EU's fundamental values and would cover aspects such as data privacy, algorithmic transparency, non-discrimination and accountability. The standards would be designed to be applicable across a range of sectors and scalable to different sizes of AI deployments, ensuring broad and flexible applicability.
- **Technical experts, with proficiencies in machine learning, data protection and ethics, would be essential** to this process. They would operate thorough evaluations of AI systems, examining technical documentation, data management practices as well as the design and

implementation processes to identify any potential risks or violations of the established standards. In cases where AI systems are used in high-stakes or sensitive contexts, more intensive audits could be conducted, possibly including elements such as stress testing and adversarial simulations to evaluate system robustness against misuse.

- For international applications, **the EU could leverage its auditing and certification mechanism as part of its foreign policy and international trade agreements,** promoting its standards as a benchmark for international cooperation in AI. The certification mark from the EU would serve as a signal to third countries and international bodies that a given AI system is trustworthy and adheres to high ethical standards. By implementing such an AI auditing and certification mechanism, the EU would reinforce its commitment to ethical AI development and use, setting a standard that could influence global practices. This would not only protect the rights of individuals within the EU but could also extend the Union's influence on the global stage, positioning it as a leader in the responsible stewardship of AI technologies.

6.2 Recommendations for the EP

The EP specifically, could reinforce and monitor the above policy processes through a variety of mechanisms. In the wake of provisional agreement on the AI Act, the EP finds itself at a pivotal juncture, tasked with navigating the delicate interplay between innovation and regulation in the realm of AI. To build effectively on this foundational legislative framework and preclude the authoritarian misuse of AI, a multi-faceted and nuanced approach is required, one that harmonises the aspirations of technological advancement with the imperatives of ethical governance.

The Parliament **could revive and render permanent its previously limited-term 'Artificial Intelligence in a Digital Age' (AIDA) Special Committee** (or a newly coined committee that specialises in AI-related issues) that would underscore the importance of continual, detailed oversight of the AI Act's implementation.

- This Committee would not only engage in **regular monitoring,** but also in **facilitating periodic reviews and legislative amendments,** particularly focusing on emerging technologies that could potentially be exploited for authoritarian purposes. Biennial reviews and hearings would become instrumental in this process, offering a platform for scrutinising compliance and adapting the legislative framework to the rapidly evolving AI landscape.

The EP is positioned to play a pivotal role in **international AI diplomacy.**

- In parallel, by **initiating dialogues and sharing regulatory best practices with legislative bodies beyond the EU, leading discussions on AI ethics at global summits, organising capacity-building with third country parliaments, making use of its influence in parliamentary fora and the Inter-Parliamentary Union,** the Parliament can extend the reach of its AI governance model. Such diplomatic engagements would be vital in advocating adoption of ethical AI standards globally, akin to the influence wielded by the GDPR in data protection. A critical component of this approach involves tightening controls over the export of AI technologies, particularly those susceptible to misuse in surveillance or repression.

The Parliament could also work towards **developing a comprehensive list of such technologies, ensuring rigorous due diligence in the export process.**

- This initiative would **necessitate close collaboration with the European Commission and the integration of stringent end-use monitoring mechanisms.** To bolster the ecosystem of ethical AI development within the EU, the Parliament could allocate specific research funds to

projects focused on creating AI systems resistant to misuse. This initiative would not only foster innovation in ethical AI development, but also support the creation of tools and methodologies for AI auditing and risk assessment.

Public awareness and education are equally crucial.

- The Parliament could launch **EU-wide campaigns to heighten public consciousness about the potential risks associated with AI misuse.** Educational seminars and workshops would further demystify AI for EU citizens, fostering an informed and engaged populace. Engaging with the private sector is another strategic axis. The Parliament could facilitate dialogues with AI developers and tech companies, encouraging adherence to ethical AI standards and promoting industry self-regulation. Recognising and endorsing ethical AI charters or agreements could significantly influence industry practices.

Moreover, **protecting whistle-blowers** who expose unethical AI practices and supporting independent media dedicated to investigating AI misuse are vital for maintaining transparency and accountability.

- The Parliament could **support legislation to safeguard these critical actors, alongside setting up funds to support investigative journalism in this domain.**

In summary, the **EP's role should be characterised by a dynamic, multi-dimensional approach.** By fostering a regulatory environment that is simultaneously adaptive, transparent and collaborative, the Parliament could ensure that AI development within the EU not only adheres to the highest ethical standards, but also sets a precedent for global AI governance. This strategy represents a concerted effort to harmonise the innovative potential of AI with the fundamental values of democracy and human rights, ensuring that AI serves as a tool for societal benefit rather than authoritarian control.

6.3 Final conclusions

Concluding an extensive discourse on EU's policy recommendations to mitigate the use of AI for repressive means, it is crucial to recognise that the trajectory of AI's application in authoritarian contexts is likely to grow in complexity and sophistication. AI technologies are evolving rapidly and their potential for misuse in enhancing state surveillance capabilities, spreading disinformation and suppressing dissent is of significant concern. The capacity for AI to analyse big data could enable authoritarian regimes to predict, pre-empt and quash potential challenges to their authority with unprecedented efficiency.

By **investing in research and development that focuses on the ethical use of AI, the EU could foster technologies** that inherently resist abusive applications and promote open, transparent AI ecosystems. The EU's role in setting a global standard for the governance of AI technologies will be instrumental. This entails not only the implementation of comprehensive legal frameworks at home but also active engagement in international fora to shape the global discourse on AI. The EU can leverage its regulatory experience with initiatives such as GDPR to advocate for international agreements that embody democratic values and human rights protections. Furthermore, the EU must also consider the dual-use nature of many AI applications, which may offer significant benefits for society, yet also possess the potential for misuse.

To address this, **the EU should refine its export controls and develop a nuanced understanding of how AI technologies can be adapted by authoritarian regimes.** A critical component will be the capacity to adjust these controls rapidly in response to technological advancements. In the realm of sanctions and accountability, the EU will need to establish mechanisms that not only penalise entities violating human rights through AI, but also provide avenues for remediation and reformation. This will involve developing a granular understanding of global supply chains and the various actors involved in the development and deployment of AI technologies. Predicting how AI may be used for repression also requires the EU to foster strong ties with academia, the private sector and civil society organisations. These partnerships could provide the early warning mechanisms and innovative solutions necessary to counteract emerging threats.

Finally, the **EU occupies a unique position to act as a 'norm-setter', leveraging its strong legal and regulatory framework to shape global AI standards.** However, it is imperative it also turns its attention inward, addressing the imminent challenges posed by AI within its borders with equal vigour. This internal focus is essential for the EU to maintain its credibility and effectiveness as a global leader in ethical AI governance. Central to the EU's strategy should be rigorous enforcement of human rights and transparency standards in AI deployment across Member States – including a closer look at how their high-technology exports are fuelling algorithmic authoritarianism abroad. This is more than a symbolic gesture; it serves as a litmus test for the EU's ability to manage AI's integration into society effectively in its own jurisdiction and its oversight mechanisms over European AI start-ups. By setting and adhering to high internal standards, the EU not only establishes a model for global AI ethics, but also ensures that these technologies are aligned with the fundamental values of democracy and human dignity within its own territory. Moreover, the EU's approach to AI regulation, exemplified by initiatives like the AI Act, needs to be implemented with a dual focus. On the one hand, it should foster innovation and technological advancement; on the other hand, it must rigorously enforce ethical compliance and conditions on European technologies contributing to repression elsewhere – and on rare occasions, within the EU itself. This balancing act is crucial in setting a precedent for how AI is developed and controlled, both within the EU and globally.

Most nations that are seeking to regulate AI comprehensively are not looking at US regulations, but at **the EU's framework as an example and a model, and the fact that the EU is indeed a global 'norm-setter' should not be taken lightly.** While the EU's role in shaping the global discourse on AI is undeniable, it is equally important for it to address the challenges within its borders with a clear and effective strategy as new AI technologies are blurring the lines between freedom of expression and security considerations every day. By doing so, the EU not only reinforces its position as a global leader in ethical AI but also ensures that the advancements in AI technology within its territory are in harmony with its commitment to human rights and the enhancement of societal welfare. It also demonstrates the courage and political will to look at itself with a more honest lens and address its shortcomings; such courage in turn encourages more honest regulations globally, and this courage is the most direct and lasting impact the EU can have on the use of AI democratically and in a way that aligns with fundamental human rights.

7 References

ABC/Reuters, 'Chinese telecom giant ZTE 'helped Venezuela develop social credit system'', ABC News, 16 November 2018.

Abozaid, A. M., 'Digital Baltaga: How Cyber Technology Has Consolidated Authoritarianism in Egypt', *SAIS Review of International Affairs*, Vol 42, No 2, 2022.

ACLU, 'MediaJustice, et al. v. Federal Bureau of Investigation, et al.', *ACLU List of FBI Cases*, 2023.

Adebahr C., and Mittelhammer, B., 'Upholding Internet Freedom as Part of the EU's Iran Policy', *Carnegie Europe*, 29 November 2023.

Agre, P. E., 'Toward a critical technical practice: Lessons learned in trying to reform AI. In Social science, technical systems, and cooperative work', *Psychology Press*, 2017.

Aizenberg, E. and van Den Hoven, J, 'Designing for human rights in AI', *Big Data & Society*, Vol 7, No 2, 2020.

Akbari, A. and Gabdulhakov, R., 'Platform surveillance and resistance in Iran and Russia: The case of Telegram', *Surveillance & Society*, Vol 17, Nos 1 and 2, 2019.

Akemi Shimoda Uechi, C. and Guimarães Moraes T., 'Brazil's path to responsible AI', 27 July 2023.

Akimenko, V. and Giles, K., 'Russia's cyber and information warfare', *Asia Policy*, Vol 15, No 2, 2020.

Atlantic Council, 'Experts react: The EU made a deal on AI rules. But can regulators move at the speed of tech?', 11 December 2023.

Al Ashry, M. S., 'A critical assessment of the impact of Egyptian laws on information access and dissemination by journalists', *Cogent Arts & Humanities*, Vol 9, No 1, 2022.

Almabdy S., and Elrefaei, L., 'Deep convolutional neural network-based approaches for face recognition', *Applied Sciences*, Vol 9, No 20, 2019.

Alterman, J. B., 'Protest, Social Media, and Censorship in Iran', *Center for Strategic and International Studies*, 18 October 2022.

Amnesty International, 'Egypt: Women influencers jailed over 'indecency': Hanin Hossam, Mawada el-Adham', 14 July 2021.

Amnesty International, 'Forensic Methodology Report: How to catch NSO Group's Pegasus', *Amnesty International*, 18 July 2021.

Amnesty International, Out of Control: Failing EU Laws for Digital Surveillance Export, 21 September 2020.

Andersen, R. S., 'Video, algorithms and security: How digital video platforms produce post-sovereign security articulations', *Security Dialogue*, Vol 48, No 4, 2017.

Andrejevic, M., Dencik, L. and Treré, E., 'From pre-emption to slowness: Assessing the contrasting temporalities of data-driven predictive policing', *New Media & Society,* Vol 22, No 9, 2020.

Anonymous Google and Amazon workers, 'We are Google and Amazon workers. We condemn Project Nimbus', *The Guardian*, 12 October 2021.

Askew, J., 'China turbocharging crackdown on Iranian women, say experts', *EuroNews*, 14 April 2023.

Backer, L. C., 'China's Social Credit System', *Current History*, Vol 118, No 809, 2019.

Bahl, V., 'No, this video doesn't show the Wagner Group moving to Belarus', *France24*, 11 July 2023.

Balkin, J. M., 'Information Fiduciaries in the Digital Age', *Balkinization*, 5 March 2014.

Ball, K., 'All consuming surveillance: surveillance as marketplace icon', *Consumption Markets & Culture*, Vol 20, No 2, 2017.

Ball, K., and Snider, L. (eds), *The surveillance-industrial complex: A political economy of surveillance,* Routledge (Oxfordshire, England, UK), 2013.

Banerjea, U., 'Revolutionary intelligence: The expanding intelligence role of the Iranian Revolutionary Guard Corps', *Journal of Strategic Security*, Vol 8, No 3, 2015.

Bazarkina, D. and Pashentsev, E., 'Malicious use of artificial intelligence', *Russia in Global Affairs*, Vol 18, No 4, 2020.

Benbouzid, B., 'To predict and to manage. Predictive policing in the United States', *Big Data & Society*, Vol 6, No 1, 2019.

Benedict, T. J., 'The Computer Got It Wrong: Facial Recognition Technology and Establishing Probable Cause to Arrest', *Wash. & Lee L. Rev.*, Vol 79, 2022.

Berger, J., 'A Dam, Small and Unsung, Is Caught Up in an Iranian Hacking Case', *The New York Times*, 25 March 2016.

Bhuihan, J., and Montgomery, B., ''A betrayal': Google workers protests Israeli military contract at vigil for ex-intern killed in airstrike', *The Guardian*, 1 December 2023.

Bhuiyan, J., 'LAPD ended predictive policing programs amid public outcry. A new effort shares many of their flaws', *The Guardian*, 8 November 2021.

Binns, R., Veale, M., Van Kleek, M. and Shadbolt, N., 'Like trainer, like bot? Inheritance of bias in algorithmic content moderation', *Social Informatics: 9th International Conference, Proceedings*, Part II 9, 2017.

Boffey, D., 'EU to launch rare inquiry into Pegasus sypware scandal', *The Guardian*, 10 February 2022.

Bolsover, G. and Howard, P., 'Computational propaganda and political big data: Moving toward a more critical research agenda', *Big data*, Vol 5, No 4, 2017.

Borak, M., 'Inside Safe City, Moscow's AI Surveillance Dystopia', *Wired*, 6 February 2023.

Borogan, I., Soldatov, A., Grossfeld, E., and Richterova, D., 'What impact has the war on Ukraine had on Russian security and intelligence?', *King's College London*, 22 February 2023.

Bouchrika, I., 'A survey of using biometrics for smart visual surveillance: Gait recognition. Surveillance in Action,' *Technologies for Civilian Military and Cyber Surveillance*, 2018.

Bozdag, E., 'Bias in algorithmic filtering and personalization. Ethics and information technology', Vol 15, 2013.

Bozorgmehr, N., 'Robots can help issue a fatwa': Iran's clerics look to harness AI', *Financial Times*, 24 September 2023

Brand L. A., 'Arab uprisings and the changing frontiers of transnational citizenship: Voting from abroad in political transitions', *Political geography*, Vol 41, 2014.

Brantingham, P. J., Valasik, M., and Mohler, G. O., 'Does predictive policing lead to biased arrests? Results from a randomized controlled trial', *Statistics and public policy*, Vol 5, No 1, 2018.

Brophy, P. and Halpin, E., 'Through the net to freedom: information, the Internet and human rights', *Journal of information science*, Vol 25, No 5, 1999.

Brown, D. K., 'Police violence and protests: Digital media maintenance of racism, protest repression, and the status quo', *Journal of Broadcasting & Electronic Media*, Vol 65, No 1, 2021.

Buchanan, B., The hacker and the state: Cyber attacks and the new normal of geopolitics, Harvard University Press, (Cambridge: United States), 2020.

Buçinca, Z., Malaya, M. B., and Gajos, K. Z., 'To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making', *Proceedings of the ACM on Human-Computer Interaction at Cornell University*, Vol 5, No CSCW1, 2021.

Burkhardt, F. and Wijermars, M., Digital Authoritarianism and Russia's War Against Ukraine: How Sanctions-induced Infrastructural Disruptions are Reshaping Russia's Repressive Capacities, *SAIS Review of International Affairs*, Vol 42, No 2, 2022.

Burton, A. M., Wilson, S., Cowan, M., and Bruce, V., 'Face recognition in poor-quality video: Evidence from security surveillance. Psychological Science', Vol 10, No 3, 2018.

Byman, D., 'The Social Media War in the Middle East', *The Middle East Journal*, Vol 75, No 3, 2021.

Caldas, R., Fadel, T., Buarque, F. and Markert, B., 'Adaptive predictive systems applied to gait analysis: A systematic review', *Gait & posture*, Vol 77, 2022.

Cameron, D., 'Docs Show FBI Pressures Cops to Keep Phone Surveillance Secrets', *Wired,* 22 June 2023.

Campbell, Z. and Chandler, C. L., 'Tools for Repression in Myanmar Expose Gap Between EU Tech Investment and Regulation', *The Intercept*, 14 June 2021.

CBS News, 'Judge limits Biden administration's contact with social media companies', 4 July 2023.

Chan, K. and Alden, C., '<Redirecting> the diaspora: China's united front work and the hyperlink networks of diasporic Chinese websites in cyberspace', *Political Research Exchange*, Vol 5, No 1, 2023.

Chen, H. and Zimbra, D., 'AI and opinion mining', IEEE Intelligent Systems, Vol 25, No 3, 2010.

Congressional Research Service, 'China's Corporate Social Credit System', IF11342, 17 January 2020.

Council of Europe, 'Council of Europe and Artificial Intelligence', Brochure, 2023.

Council of Europe Revised Zero Draft [Framework] Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law, Committee of Artificial Intelligence, 6 January 2023.

Council of Europe, Algorithms and Human Rights: Study on the human rights dimensions of automated data processing techniques and possible regulatory implications, 2018.

Council of Europe, CEPEJ European Ethical Charter on the use of artificial intelligence (AI) in judicial systems and their environment, 2024.

Crawford, K., *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*, Yale University Press (New Haven: Connecticut, USA), 2021.

Creemers, R., 'China's Social Credit System: an evolving practice of control', 2018.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A., 'Generative adversarial networks: An overview', *IEEE signal processing magazine*, Vol 35, No 1, 2018.

Crosston, M., 'Cyber colonization: The Dangerous Fusion of Artificial Intelligence and Authoritarian Regimes', *Cyber, Intelligence, and Security Journal*, Vol 4, No 1, 2020.

Curtis, M., 'Aclu Challenges Use Of "Stingray" Surveillance Technology By Baltimore Police', *ACLU Maryland*, 26 November, 2014.

Daniels, J. and Murgia, M., 'Deepfake 'news' videos ramp up misinformation in Venezuela', *Financial Times*, 17 March 2023.

Darer, A., Farnan, O. and Wright, J., 'FilteredWeb: A framework for the automated search-based discovery of blocked URLs', Network Traffic Measurement and Analysis Conference, *Institute of Electrical and Electronics Engineers*, 2017.

Darwish, O, Tashtoush, Y., Maabreh, M., Al-essa, R., Aln'uman, R., Alqublan, A., and Elkhodr, M., 'Identifying Fake News in the Russian-Ukrainian Conflict Using Machine Learning', *International Conference on Advanced Information Networking and Applications*, Vol 665, 2023.

Davenport, C., *State repression and the domestic democratic peace*, Cambridge University Press (Cambridge: United Kingdom), 2007.

Deloitte, 'The China Personal Information Protection Law (PIPL)', May 2021.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 'Bert: Pre-training of deep bidirectional transformers for language understanding', ACL Anthology, 2018.

Dias, F. D. V. and Amaral, A. J. do, 'Media criminology in Brazil: algorithms and new geopolitic control developments' *Revista Brasileira Estudos Politicos*, Vol 123, 2021.

Dirbaba, B. O. and O'Donnell, P., 'The double talk of manipulative liberalism in Ethiopia: An example of new strategies of media repression', *African Communication Research*, Vol 5, No 3, 2012.

Dixon, P., 'A Failure to "Do No Harm" -- India's Aadhaar biometric ID program and its inability to protect privacy in relation to measures in Europe and the U.S.', *Health Technol (Berl)*, National Institutes of Health, 2017.

Dragu, T. and Lupu, Y., 'Digital authoritarianism and the future of human rights', *International Organization*, Vol 75, No 4, 2021.

Earl, J., Mahe, R. T. V. and Pan, J., 'The digital repression of social movements, protest, and activism: A synthetic review', *Science Advances*, Vol 8, No 10, 2022.

Edel, M. and Josua, M., 'How authoritarian rulers seek to legitimize repression: framing mass killings in Egypt and Uzbekistan', *Democratization*, Vol 25, No 5, 2018.

Egbert, S., 'About discursive storylines and techno-fixes: the political framing of the implementation of predictive policing in Germany', *European Journal for Security Research*, Vol 3, 2018.

Egbert, S., 'Predictive policing and the platformization of police work', *Surveillance & Society*, Vol 17, No 1 and 2, 2019.

Ellis, R. E., 'China's Economic Struggle for Position in Latin America', *China Engages Latin America: Distorting Development and Democracy?*, 2022.

El-Maghraby, R. T., Abd Elazim, N. and Bahaa-Eldin, A., 'A survey on deep packet inspection', 12th International Conference on Computer Engineering and Systems, 2017.

Elswah, M. and Alimardani, M., 'Propaganda Chimera: Unpacking the Iranian Perception Information Operations in the Arab World', *Open Information Science*, Vol 5, No 1, 2021.

Engstrom, D. F., Ho, D. E., Sharkey, C. M., and Cuéllar, M. F., 'Government by algorithm: Artificial intelligence in federal administrative agencies', *NYU School of Law, Public Law Research Paper*, 2020.

Ermoshina, K., Loveluckv, B., and Musiani, F., 'A market of black boxes: The political economy of Internet surveillance and censorship in Russia', *Journal of Information Technology & Politics,* Vol 19, No 1, 2022.

Eslami, M., Mosavi, N. S. and Can, M., 'Sino-Iranian cooperation in artificial intelligence: A potential countering against the US Hegemony', *The Palgrave Handbook of Globalization with Chinese Characteristics: The Case of the Belt and Road Initiative,* 2023.

Espinoza, J., 'EU competition chief defends Artificial Intelligence Act after Macron's attack', *Financial Times*, 29 December 2023.

Ettlinger, N., 'Algorithmic affordances for productive resistance', *Big Data & Society*, Vol 5, No 1, 2018.

Eubanks, V., Automating inequality: How high-tech tools profile, police, and punish the poor, St. Martin's Press (New York, New York State: United States of America), 2018.

European Commission, Communication from the Commission to the European Parliament and the Council: Advancing European economic security: an introduction to five new initiatives, COM(2024) 22 final, 24 January 2024.

European Commission, 'Strengthened EU export control rules kick in', Press Release, 9 September 2021.

European Commission, Proposal for a regulatıon of the European Parlıament and of the Council layıng down harmonısed rules on artıfıcıal ıntellıgence (artıfıcıal ıntellıgence act) and amendıng certaın unıon legıslatıve acts, COM(2021) 206 final, 21 April 2021.

European Commission, Study on the impact of new technologies on free and fair elections, DG JUST Election Study, March 2021.

European 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, 24 January 2021.

European Commission, 'Ethic guidelines for trustworhty AI', 8 April 2019.

European Council, 'Artificial intelligence act: Council and Parliament strike a deal on the first rules for AI in the world', 9 December 2023.

European Data Protection Board, Guidelines 05/2022 on the use of facial recognition technology in the area of law enforcement, Version 1.0, 2022.

European External Action Service, '1st EEAS Report on Foreign Information Manipulation and Interference Threats', 7 February 2023.

European Parliament Draft Recommendation to the Council and the Commission following the investigation of alleged contraventions and maladministration in the application of Union law in relation to the use of Pegasus and equivalent surveillance spyware, (2023/2500(RSP)), 22 May 2023.

European Parliament, Draft Report on the proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts, 2021/0106(COD), 22 May 2023.

European Parliament, 'IN-DEPTH ANALYSIS for the PEGASUS committee, Pegasus and surveillance spyware', Policy Department for Citizens' Rights and Constitutional Affairs Directorate-General for Internal Policies, PE 732.268, 2022.

European Parliament, Biometric Recognition and Behavioural Detection, Briefing for the JURI and PETI committees, PE 697.131, September 2021.

Fang, H. and Qian, Q.,' Privacy preserving machine learning with homomorphic encryption and federated learning', *Future Internet*, Vol 13, No 4, 2021.

Faucon, B., 'U.A.E. Trade Provides Iran With Western Goods, From Perfume to Laptops', *The Wall Street Journal*, 5 July 2022.

Feldstein S. and Kot, B., 'Why Does the Global Spyware Industry Continue to Thrive? Trends, Explanations, and Responses', Carnegie, 14 March 2023.

Feldstein, S. and Youngs, R., 'Pegasus and the EU's external relations', Policy Department for Citizens' Rights and Constitutional Affairs Directorate-General for Internal Policies, PE 741.475, 2023.

Feldstein, S. *The rise of digital repression: How technology is reshaping power, politics, and resistance*, Oxford University Press (Oxford: United Kingdom), 2021.

Feldstein, S., 'The Road to Digital Unfreedom: How Artificial Intelligence Is Reshaping Repression', *Journal of Democracy*, Vol 30, No 1, 2019.

Feliba, D., 'How AI shaped Milei's path to Argentina presidency', *The Japan Times*, 22 November 2023.

Ferri, P., 'Witness in Pegasus case accuses Peña Nieto of ordering spying operation on Carlos Slim', *El Pais,* 5 December 2023.

Flaherty, D. H., 'The emergence of surveillance societies in the Western world: Toward the year 2000', *Government Information Quarterly*, 1988.

Fuchs, C., 'Societal and ideological impacts of deep packet inspection internet surveillance', *Information, Communication & Society*, Vol 16, No 8, 2013.

Funk, A., Shahbaz, A. and Veseinsson, K., 'Freedom on the Net 2023: The Repressive Power of Artificial Intelligence', Freedom House, 2023.

Gallagher, R., 'Russia-Linked Hackers Claim Credit for OpenAI Outage This Week', *Bloomberg*, 9 November 2023.

Galperin, E., 'Swedish Telcom Giant Teliasonera Caught Helping Authoritarian Regimes Spy on Their Citizens', *Electronic Frontier Foundation*, 18 May 2012.

Gaufman, E., 'Cybercrime and Punishment: Security, Information War, and the Future of Runet', in Gritsenko D., Wijermas, M. and Kopotev, M. (eds), *The Palgrave Handbook of Digital Russia Studies,* 2021, pp. 115-134.

Gee, H., 'Almost Gone: The Vanishing Fourth Amendment's Allowance of Stingray Surveillance in a Post-Carpenter Age', *The Southern California Review of Law and Social Justice,* Vol 28, 2019.

Geneva Internet Platform Dig Watch, 'France and partners propose a programme of action for advancing responsible state behaviour in cyberspace', 8 October 2020.

Ghiabi ,M., 'The council of expediency: crisis and statecraft in Iran and beyond', *Middle Eastern Studies*, 2019, Vol 55, No 5.

Głowacka, D., Youngs, R., Pintea, A. and Wołosik, E., 'Digital technologies as a means of repression and social control', Policy Department for External Relations, PE 653.636, 2021.

Gohdes, A. R., 'Repression technology: Internet accessibility and state violence', *American Journal of Political Science*, Vol 64 No 3, 2020.

Gohdes, A. R., 'Studying the Internet and Violent Conflict', *Conflict Management and Peace Science*, Vol 5, No 1, 2018.

Goldstein, R. J., *Political repression in modern America from 1870 to 1976*, University of Illinois Press (Champaign, Illinois: United States of America), 2001.

Golovchenko, Y., 'Fighting propaganda with censorship: A study of the Ukrainian ban on Russian social media', *The Journal of Politics*, Vol 84, No 2, 2022.

Gravett, W. H., 'Digital neocolonialism: the Chinese surveillance state in Africa', *African Journal of International and Comparative Law*, Vol 30, No 1, 2022.

Gray, M., 'Urban surveillance and panopticism: will we recognize the facial recognition society?', *Surveillance & Society*, Vol 1, No 3, 2003.

Greenberg, A., 'Hacking Team Breach Shows a Global Spying Firm Run Amok', 6 July 2015.

Greitens, S. C. 'Dealing with demand for China's global surveillance exports', *Brookings Institute*, April 2020.

Greitens, S. C., Lee, M., and Yazici, E., 'Counterterrorism and preventive repression: China's changing strategy in Xinjiang', *Harvard Kennedy School of International Affairs International Security Quarterly*, Vol 44, No 3, 2019.

Grinberg, D., 'Chilling developments: digital access, surveillance, and the authoritarian dilemma in Ethiopia', *Surveillance & Society*, Vol 15, No 3/4, 2017.

Grinko, M., Qalandar, S., Randall, D., and Wulf, V., 'Nationalizing the Internet to Break a Protest Movement: Internet Shutdown and Counter-Appropriation in Iran of Late 2019', *Proceedings of the ACM on Human-Computer Interaction*, Vol 6, No CSCW2, 2022.

Grön, K., Chen, Z., and Ruckenstein, M., 'Concerns with Infrastructuring: Invisible and Invasive Forces of Digital Platforms in Hangzhou, China', *International Journal of Communication*, Vol 17, 2023.

Guggenberger, N., and Salib, P. N., 'From Fake News to Fake Views: New Challenges Posed by ChatGPT-Like AI', Lawfare Institute, *The Brookings Institution*, 20 January 2023.

Gurkov, A, 'Personal Data Protection in Russia', *The Palgrave Handbook of Digital Russia* Studies, 2021.

Gurr, T. G., 'War, revolution, and the growth of the coercive state', *Comparative Political Studies*, Vol 21, No 1, 1998.

Hardyns, W., and Rummens, A., 'Predictive policing as a new tool for law enforcement? Recent developments and challenges', *European journal on criminal policy and Research*, Vol 24, 2018.

Harris, E. J., Khoo, I. H., and Demircan, E., 'A survey of human gait-based artificial intelligence applications', *Frontiers in Robotics and AI,* Vol 8, 2022.

Harris, R., 'Repression and quiet resistance in Xinjiang' *Current History*, Vol 118, No 810, 2019.

Hassib, B. and Shires, J., 'Manipulating uncertainty: cybersecurity politics in Egypt', *Journal of Cybersecurity*, Vol 7, No 1, 2021.

Hawkins, S., 'Cell-Site Clone to Track Narcotics Suspect Approved, With Limits,' *Bloomberg Law*, 25 August 2022.

Heickero, R., 'Russia's information warfare capabilities. In Current and Emerging Trends in Cyber Operations: Policy, Strategy and Practice', *Palgrave Studies in Cybercrime and Cybersecurity*, 2015.

Herlocker, J., Konstan, A. and Riedl, J., 'An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms', *Information retrieval*, Vol 5, 2002.

Hill, K., 'Wrongfully accused by an algorithm', *In Ethics of Data and Analytics, Auerbach Publications*, 2022.

Hillman, J., Sacks, D., Lew, J. J, and Roughead, G., 'China's Belt and Road: Implications for the United States', *New York: Council on Foreign Relations*, 2021.

Hobbs, W. R. and Roberts, M. E., 'How sudden censorship can increase access to information', *American Political Science Review*, Vol 112, No 3, 2018.

Hogue, S., 'Civilian Surveillance in the War in Ukraine: Mobilizing the Agency of the Observers of War', *Surveillance & Society*, Vol 21, No 1, 2023.

Hou, R, 'Neoliberal governance or dtigitalized autocracy? The rising market for online opinion surveillance in China', *Surveillance & Society,* Vol 15, No 3 and 4, 2017.

Howard, P. N., The digital origins of dictatorship and democracy: Information technology and political Islam, Oxford University Press (Oxford: United Kingdom), 2010.

Howell, O'Neill P., 'French spyware bosses indicted for their role in the torture of dissidents', *MIT Technology Review*, 22 June 2021.

Hsu, T., and Myers, S. L., 'Pro-China YouTube Network Used A.I. to Malign U.S., Report Finds', *The New York Times*, 14 December 2023.

Human Rights Watch, 'Unprecedented Repression Demands Unprecedented Response', 3 October, 2023.

Hurcombe, L., Yong Neo, H. and Wong, D., 'China's cyberspace regulator releases draft measures for managing generative AI services', DLA Piper, *Loxology*, 18 April 2023.

IEEE SA, 'Active Standard: IEEE 7000-2021, IEEE Standard Model Process for Addressing Ethical Concerns during System Design', 15 September 2021.

IEEE SA, 'Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems', Version 2, 2019.

Iliadis, A. and Acker, A., 'The seer and the seen: Surveying Palantir's surveillance platform', *The Information Society*, Vol 38, No 5, 2022.

India government, National Strategy AI For All, June 2018.

Indiacode, 'The Information Technology Act, 2000', 2000.

Innes de Neufville, J., 'Human rights reporting as a policy tool: An examination of the State Department Country Reports', *Human Rights Quarterly*, Vol 8, 1986.

Insikt Group, 'Obfuscation and AI Content in the Russian Influence Network "Doppelgänger" Signals Evolving Tactics', *Recorded Future – Russia Threat Analysis*, 5 December 2023.

International Federation for Human Rights, 'Surveillance and torture in Egypt and Libya: Amesys and Nexa Technologies executives indicted', Press release, 22 June 2021.

Introna, L. and Wood, D., 'Picturing algorithmic surveillance: The politics of facial recognition systems', *Surveillance & Society*, Vol 2, No 2/3, 2004.

Ioannou, D. 'Deepfakes, Cheapfakes, and Twitter Censorship Mar Turkey's Elections', *Wired*, 26 May 2023.

Jamil, S., 'Automated journalism and the freedom of media: Understanding legal and ethical implications in competitive authoritarian regime', *Journalism Practice*, Vol 17, No 6, 2023.

Japan government, 'Annex. G20 AI Principles 1. The G20 supports the Principles for responsible stewardship of Trustworthy AI in Section 1' and takes note of the Recommendations in Section 2', 2019.

Jiang F., and Xie, C., 'Roles of Chinese police amidst the COVID-19 pandemic', *Policing: A Journal of Policy and Practice*, Vol 14, No 4.

Johnson, W. J., and Valente, A., 'Tactical language and culture training systems: Using AI to teach foreign languages and cultures', *AI magazine*, Vol 30, No 2, 2009.

Jones, E., 'Digital disruption: artificial intelligence and international trade policy', *Oxford Review of Economic Policy*, Vol 39, No 1, Spring 2023.

Jones, J., 'Deepfake of purported Putin declaring martial law fits disturbing pattern', *MSNBC,* 7 June 2023.

Jones, M. O., Digital authoritarianism in the Middle East: Deception, disinformation and social media, Hurst Publishers (London: United Kingdom), 2022.

Josua, M. and Edel, M., 'The Arab uprisings and the return of repression', *Mediterranean Politics*, Vol 26, No 5, 2021.

Juneau, T., 'Iran's policy towards the Houthis in Yemen: a limited return on a modest investment', *International Affairs*, Vol 92, No 3, 2016.

Kam S. and Clarke M., 'Securitization, surveillance and 'De-extremization' in Xinjiang', *International Affairs*, Vol 97, No 3, 2021.

Kaneko, T., Kameoka, H., Tanaka, K., and Hojo, N., 'Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion', *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP*), 2019.

Kanetake, M., 'The EU's export control of cyber surveillance technology: human rights approaches', *Business and Human Rights Journal*, Vol 4 No 1, 2019.

Kang, D., 'Chinese 'gait recognition' tech IDs people by how they walk', *Associated Press*, 6 November 2018.

Kargar, S. and Rauchfleisch, A., 'State-aligned trolling in Iran and the double-edged affordances of Instagram', *New media & society*, Vol 21, No 7, 2019.

Kaster, S. D., and Ensign, P. C., 'Privatized espionage: NSO Group Technologies and its Pegasus spyware', *Thunderbird International Business Review,* Vol 65, No 3, 2023.

Kaufmann, M., Egbert, S., and Leese, M., 'Predictive policing and the politics of patterns', *The British Journal of criminology*, Vol 59, No 3, 2019.

Keating, J., 'Why the U.S. Government Took Down Dozens of Iranian Websites This Week', *Slate*, 24 June 2021.

Keegan, M., 'Big Brother is watching: Chinese city with 2.6m cameras is world's most heavily surveilled', *The Guardian*, 2 December 2019.

Keremoğlu, E., and Weidmann, N. B., 'How dictators control the internet: A review essay', *Comparative Political Studies*, Vol 53 No 10-11, 2020.

Kerr, S., and Bozorgmehr, N., 'UAE boosts trade with Iran after eased restrictions on business activity', *Financial Times*, 10 September 2023.

Kharazi, A., 'Authoritarian Surveillance: 'A Corona Test', *Surveillance and Society*, Vol 19, No 1, 2021.

King, G., Pan, J. and Roberts, M. E., 'Reverse-engineering censorship in China: Randomized experimentation and participant observation', *Science*, Vol 345, No 6199, 2014.

King, G., Pan, JI and Roberts, M. E, 'How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument', *American Political Science Review*, Vol 111, No, 2017.

King-Wa F., 'Propagandization of Relative Gratification: How Chinese State Media Portray the International Pandemic', *Political Communication,* Vol 40, No 6.

Kirchgaessner, S., 'Israeli spyware company NSO Group placed on US blacklist', *The Guardian*, 3 November 2021.

Kirchgaessner, S., 'More Polish opposition figures found to have been targeted by Pegasus spyware,' *The Guardian*, 17 February 2022.

Kitroeff, N., and Bergman, R., 'Why Did a Drug Gang Kill 43 Students? Text Messages Hold Clues.', The New York Times, 2 September 2023.

Klingert, L., 'Belgian police reveal use of controversial Pegasus spyware', *The Brussels Times*, 21 April 2022.

Kostka, G., Steinacker L., and Meckel, M., 'Under big brother's watchful eye: Cross-country attitudes toward facial recognition technology', *Government Information Quarterly*, Vol 40, No 8, 2023.

Kotliar, D. M. and Carmi, E., 'Keeping Pegasus on the wing: legitimizing cyber espionage', *Information, Communication & Society*, 2023.

Kumar, V. D. A., Malathi, S., Vengatesan K., and Ramakrishnan, M., 'Facial recognition system for suspect identification using a surveillance camera', *Pattern Recognition and Image Analysis*, Vol 28, 2018.

Laird, B., 'The Risks of Autonomous Weapons Systems for Crisis Stability and Conflict Escalation in Future U.S.-Russia Confrontations', *RAND Corporation*, 3 June 2020.

Lamoreaux, J. W. and Flake, L., 'The Russian Orthodox Church, the Kremlin, and religious (il) liberalism in Russia', *Palgrave Communications*, Vol 4, No 1, 2018.

Landman, T., 'Holding the line: Human rights defenders in the age of terror', *The British Journal of Politics and International Relations*, Vol 8, No 2, 2006.

Larsson, S., 'On the governance of artificial intelligence through ethics guidelines', *Asian Journal of Law and Society*, Vol 7, No 3, 2020.

Le Monde, 'Macron argues against 'punitive' AI regulation', 17 November 2023.

Lee, T. K., Belkhatir, M., and Sanei, S., 'A comprehensive review of past and present vision-based techniques for gait recognition', *Multimedia tools and applications,* Vol 72, 2014.

Lee-Wee, S. and Mozur, P., 'China Uses DNA to Map Faces, With Help From the West', *The New York Times*, 3 December 2023.

Leibold, J., 'Surveillance in China's Xinjiang region: Ethnic sorting, coercion, and inducement', *Journal of contemporary China*, Vol 29, No 121, 2020.

Leloup, P. D. and Utersinger, M., '« Projet Pegasus » : un téléphone portable d'Emmanuel Macron dans le viseur du Maroc', *Le Monde,* 20 July 2021.

Levitt, M., 'Iran's Deadly Diplomats', *CTC Sentinel*, Vol 16, 2018.

Liang, F., Das V., Kostyuk, N. and Hussain, M. M., 'Constructing a data-driven society: China's social credit system as a state surveillance infrastructure', *Policy & Internet*, Vol 10, No 4, 2018.

Liu, C., 'Multiple social credit systems in China', *Economic Sociology: The European Electronic Newsletter*, Vol 21, No 1, 2019.

Lyngaas, S., 'Meta identifies Chinese propaganda threat ahead of 2024 election', *CNN Business*, 30 November 2023.

Lyu, X., Chen, Z., Wu, D. and Wang, W., 'Sentiment analysis on Chinese Weibo regarding COVID-19', Proceedings, Natural Language Processing and Chinese Computing: 9th CCF International Conference, 2020.

Mac Dougall, D., 'Spies like us: How does Russia's intelligence network operate across Europe?', *EuroNews*, 18 August 2023.

Mac, Síthigh D. and Siems, M., 'The Chinese social credit system: A model for other countries?', *The Modern Law Review,* Vol 82, No 6, 2019.

Madiega, T., 'European Parliament Briefing on EU Legislation on Artificial Intelligence', PE 698.792 – June 2023.

Maity, S., Abdel-Mottaleb, M., and Asfour, S. S., 'Multimodal low resolution face and frontal gait recognition from surveillance video', *Electronics*, Vol 10, No 9, 2021, p. 1013.

Makihara, Y., Nixon, M. S., and Yagi, Y., 'Gait recognition: Databases, representations, and applications', *Computer Vision: A Reference Guide*, 2020.

Mane, S. and Shah, G., 'Facial recognition, expression recognition, and gender identification', *In Data Management, Analytics and Innovation: Proceedings of ICDMAI 2018*, Vol 1, 2019.

Marchant de Abreu, C., 'These images don't show confrontations between Wagner group and Russian army', *France24*, 26 June 2023.

Marcus J. S., Poitiers, N., De Ridder, M., and Weil, P., 'The decoupling of Russia: high-tech goods and components', *Bruegel*, 28 March 2022.

Marczak, B., Scott-Railton, J., and Deibert, R., 'NSO Group Infrastructure Linked to Targeting of Amnesty International and Saudi Dissident', *The Citizen Lab,* 31 July 2018.

Maréchal, N., 'Networked authoritarianism and the geopolitics of information: Understanding Russian Internet policy', *Media and communication*, Vol 5, No 1, 2017.

Marelli, M., 'The SolarWinds hack: Lessons for international humanitarian organizations', *International Review of the Red Cross*, Vol 104, No 919, 2022.

Margolis, M. and Muggah R., 'Brazil's fake-news problem won't be solved before Sunday's vote', 27 October 2022.

Martin, F., 'Overseas study as zone of suspension: Chinese students re-negotiating youth, gender, and intimacy. Journal of Intercultural Studies', Vol 39, No 6, 2018.

Masri, L., 'Facial recognition is helping Putin curb dissent with the aid of U.S. tech', *Reuters*, 28 March 2023.

Mazzeti, M., Bergman, R., and Gridneff, M. S., 'How the Global Spyware Industry Spiraled Out of Control', *The New York Times*, 8 December 2022.

McSorley, T., 'The Case for a Ban on Facial Recognition Surveillance in Canada', Surveillance & Society, Vol 19, No 2, 2021.

Meijer A., and Wessels M., 'Predictive policing: Review of benefits and drawbacks', *International Journal of Public Administration*, Vol 42, No 12, 2019, pp.1031-1039.

Mejias, U. A. and Vokuev, N. E., 'Disinformation and the media: the case of Russia and Ukraine', *Media, culture & society*, Vol 39, No 7, 2017.

Michaelsen, M., 'Authoritarian practices in the digital age| transforming threats to power: The International Politics of Authoritarian Internet Control in Iran', *International Journal of Communication*, Vol 12, 2018.

Michaelsen, M., 'Far away, so close: Transnational activism, digital surveillance and authoritarian control in Iran', *Surveillance & Society*, Vol 15, Nos 3 and 4, 2017.

Mir, U. B., Kar, A. K., Dwivedi, Y.K., Gupta, M. P. and Sharma, R. S., 'Realizing digital identity in government: Prioritizing design and implementation objectives for Aadhaar in India', *Government Information Quarterly*, Vol 37, No 2, 2021.

Mittelstadt, B., 'Automation, algorithms, and politics| auditing for transparency in content personalization systems', *International Journal of Communication*, No 10, 2016, p. 12.

Mökander, J., Juneja, P., Watson, D. S., and Floridi, L., 'The US Algorithmic Accountability Act of 2022 vs. The EU Artificial Intelligence Act: what can they learn from each other?', *Minds and Machines*, Vol 32, No 4, 2022.

Moss, D. M., Michaelsen, M. and Kennedy, G., 'Going after the family: Transnational repression and the proxy punishment of Middle Eastern diasporas', *Global Networks*, Vol 22, No 4, 2022.

Mouloodi, S., Rahmanpanah, H., Gohari, S., Burvill, C., Tse, K. M. and Davies, H. M., 'What can artificial intelligence and machine learning tell us? A review of applications to equine biomechanical research', *Journal of the Mechanical Behavior of Biomedical Materials*, Vol 123, 2021.

Moyakine, E. and Tabachnik, A., 'Struggling to strike the right balance between interests at stake: The 'Yarovaya', 'Fake news' and 'Disrespect'laws as examples of ill-conceived legislation in the age of modern technology', *Computer Law & Security Review*, Vol 40, 2021.

Mozur, P., 'Inside China's Dystopian Dreams: A.I., Shame and Lots of Cameras', *The New York Times*, 8 July 2018.

Munoriyarwa, A., 'The militarization of digital surveillance in post-coup Zimbabwe:'Just don't tell them what we do', *Security Dialogue,* Vol 53, No 5, 2022.

Namdarian, L., Alidousti, S., and Rasuli, B., 'Developing a comprehensive framework for analyzing national scientific and technical information policy: application of HeLICAM in Iran', *Online Information Review*, Vol 45, No 7, 2021.

Nechushtai, E., Zamith, R., and Lewis, S. C., 'More of the Same? Homogenization in News Recommendations When Users Search on Google, YouTube, Facebook, and Twitter', *Mass Communication and Society*, 2023.

Neugeboren, E., 'Pegasus Spyware Targets Moroccan Journalist', *Voice of America*, 26 June 2020.

Newman, L. H., 'Instagram Slow to Tackle Bots Targeting Iranian Women's Groups', *Wired*, 19 July 2022.

Nikkarila J. P., and Ristolainen, M., 'RuNet 2020'-Deploying traditional elements of combat power in cyberspace?' in *2017 International Conference on Military Communications and Information Systems*, 15-16 May 2017.

Nkonde, M., 'Automated anti-blackness: facial recognition in Brooklyn, New York', *Harvard Journal of African American Public Policy*, Vol 20, 2019.

Noble, S. U., *Algorithms of oppression: How Search Engines Reinforce Racism*. New York University Press (Manhattan, New York State: USA), 2018.

Novaya Gazeta Europe, 'Roskomnadzor plans to use AI to monitor 'manipulations and social polarisation' online', *Novaya Gazeta Europe*, 8 February 2023.

Observatory of Economic Complexity, 'Economic Complexity Indicators of the Islamic Republic of Iran', 2021-2022.

Office of the President of the USA, 'Charter of the Machine Learning and Artificial Intelligence, Committee on Technology, National and Science Technology Council', 2016.

Ogasawara, M., 'Mainstreaming colonial experiences in surveillance studies', *Surveillance & Society*, Vol 17, No 5, 2019.

Olcott, E., and S. Yu, S., 'China escalates zero-Covid propaganda effort as experts warn of economic damage', *The Financial Times*, 14 April 2022.

One Trust Data Guidance, 'Kenya: Bill on Robotics and AI society introduced to National Assembly', 4 December 2023.

O'Neil, C., *Weapons of math destruction: How big data increases inequality and threatens democracy*, Penguin Random House LLC (New York, New York State: USA), 2017.

Oztig, L., 'Big data-mediated repression: a novel form of preemptive repression in China's Xinjiang region', *Contemporary Politics*, 2023.

Pan, J. and Siegel, A. A., How Saudi crackdowns fail to silence online dissent, *American Political Science Review*, Vol 114, No 1, 2020.

Pan, J., 'How Market Dynamics of Domestic and Foreign Social Media Firms Shape Strategies of Internet Censorship', *Problems of Post-Communism*, Vol 64, No 3, 2017.

Parkin, B., 'Deepfakes for $24 a month: how AI is disrupting Bangladesh's election', *Financial Times*, 14 December 2023.

Pascu, L., 'Russian Ministry testing gait recognition as part of national biometric surveillance system', BiometricUpdate.com, 25 February 2020.

Pasquale, F., The black box society: The secret algorithms that control money and information, Harvard University Press (Cambridge, Massachusetts: USA), 2015.

Paul, M. L., 'Noah' and 'Daren' report good news about Venezuela. They're deepfakes', *The Washington Post*, 2 March 2023.

Petrella, S., Miller, C. and Cooper, B., 'Russia's artificial intelligence strategy: the role of state-owned firms', *Orbis*, Vol 65, No 1, 2021.

Pisanu G., and Arroyo, V., 'Made Abroad, Deployed at Home', AccessNow, 2021.

Powell, B., Avidan, E. and Latifi, S., 'Threat Recognition from Gait Analysis', *IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference*, 2019.

Prabowo, R. and Thelwall, M., 'Sentiment analysis: A combined approach', *Journal of Informetrics*, Vol 3, No 2, 2009.

Pratt, N. and Rezk, D., 'Securitizing the Muslim Brotherhood: State violence and authoritarianism in Egypt after the Arab Spring', *Security Dialogue*, Vol 50, No 3, 2019.

Priest, D., 'A UAE agency put Pegasus spyware on phone of Jamal Khashoggi's wife months before his murder, new forensics show', *The Washington Post*, 21 December 2021.

Ragas, J., 'A starving revolution: ID cards and food rationing in Bolivarian Venezuela', *Surveillance & Society*, Vol 15, Nos 3 and 4, 2017.

Rahimi, B., 'The politics of the Internet in Iran', *Media, culture and society in Iran*, 2007.

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., and Denton, E., 'Saving face: Investigating the ethical concerns of facial recognition auditing', *In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society,* 2020.

Ramesh, R., Raman, R. S., Virkud, A., Dirksen, A., Huremagic, A., Fifield, D. and Ensafi, R., 'Network responses to Russia's invasion of Ukraine in 2022: a cautionary tale for internet freedom', *32nd USENIX Security Symposium*, USENIX Security, Vol 23, 2023.

Rankin, J., 'EU urged to tighten spyware safeguards in wake of Pegasus revelations', *The Guardian*, 9 May 2023.

Raso, F., Hilligoss, H., Krishnamurthy, V., Bavitz, C., and Kim, L., 'Artificial intelligence & human rights: Opportunities & risks', *Berkman Klein Center Research Publication*, 2018.

Ravi, K., and Ravi, V., 'A survey on opinion mining and sentiment analysis: tasks, approaches and applications', *Knowledge-based systems*, Vol 89, 2015.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Official Journal of the European Union, 119, 4 May 2016.

Rejali, D., *Torture and democracy*, Princeton University Press (Princeton, New Jersey, United States of America), 2009.

Reuters Staff, 'Dutch recall ambassador to Iran after diplomats expelled', *Reuters*, 4 March 2029.

Reuters, 'Russia plans to try to block VPN services in 2024 – senator', *Reuters*, 3 October 2023.

Rezende, I. N., 'Facial recognition in police hands: Assessing the 'Clearview case' from a European perspective', *New Journal of European Criminal Law*, Vol 11, No 3, 2020.

Richardson, R., Schultz, J. M. and Crawford, K., 'Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems', and justice, *NYUL Rev. Online*, 2019.

Rober,s, H., Cowls, J., Morley, J., Taddeo, M., Wang, V. and Floridi, L., 'The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation', *AI & Society*, Vol 36, 2020.

Roberts, M. E., 'Resilience to online censorship', Annual Review of Political Science, Vol 23.

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V. and Floridi, L., 'The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation', *AI & Society*, Vol 36, 2020.

Roberts, M., *Censored: distraction and diversion inside China's Great Firewall*, Princeton University Press (Princeton, New Jersey, USA), 2018.

Roberts, S. R., 'The biopolitics of China's "war on terror" and the exclusion of the Uyghurs', *Critical Asian Studies*, Vol 50, No 2, 2018.

Roberts, T. (ed.), Digital Rights in Closing Civic Space: Lessons from Ten African Countries, Institute of Development Studies, 2021.

Robinson, O., Robinson Z., and Sardarizadeh, S., 'Ukraine war: How TikTok fakes pushed Russian lies to millions', *BBC News*, 15 December 2023.

Roche, G. and Leibold, J., 'State Racism and Surveillance in Xinjiang (People's Republic of China)', *The Political Quarterly*, Vol 93, No 3, 2022.

Rød E. G., Rustemeyer, J. and Otto, S., Introducing the MMAD Repressive Actors Dataset, *Research & Politics*, Vol 10, No 2, 2023.

Roussi, A., 'How Europe became the Wild West of spyware', *Politico*, 25 October 2023.

Rudie, J. D., Katz, Z., Kuhbander, S., and Bhunia, S., 'Technical Analysis of the NSO Group's Pegasus Spyware,' IEE Explore', 2021.

Ryan-Mosley, T., 'This huge Chinese company is selling video surveillance systems to Iran', *MIT Technology Review*, 15 December 2021.

Saheb, T., 'Ethically contentious aspects of artificial intelligence surveillance: a social science perspective', *AI and Ethics*, Vol 3, No 2, 2023.

Sanger, D. E., Perlroth, N., Swanson, A., and Bergman, R., 'U.S. Blacklists Israeli Firm NSO Group Spyware', *The New York Times*, 3 November 2021.

Saudi government, 'Summary of discussions from the G20 AI dialogue in 2020', 2020.

Saunders, J., Hunt, P., and Hollywood, J. S., 'Predictions put into practice: a quasi-experimental evaluation of Chicago's predictive policing pilot', *Journal of Experimental Criminology,* Vol 12, 2016.

Sayed, N., 'Towards the Egyptian revolution: Activists' perceptions of social media for mobilization', *Journal of Arab & Muslim Media Research*, Vol 4, No 2-3, 2012.

Schlumberger, O., Edel, M., Maati, A. and Saglam, K., 'How Authoritarianism Transforms: A Framework for the Study of Digital Dictatorship', *Government and Opposition*, 2023.

Scott-Railton, J., Campo, E., Marczak, B., Razzak, B. A., Anstis, S., Böcü, G. and Deibert, R., 'Catalangate: Extensive mercenary spyware operation against Catalans using pegasus and Candiru', *The Citizen Lab*, 2022.

Seah, C. W., Chieu, H. L., Chai, K. M. A., Teow, L. N. and Yeong, L. W., 'Troll detection by domain-adapting sentiment analysis', 18th International Conference on Information Fusion, Institute of Electrical and Electronics Engineers, 2015.

Shahi, A. and Abdoh-Tabrizi, E., 'Iran's 2019–2020 demonstrations: the changing dynamics of political protests in Iran', *Asian Affairs*, Vol 51, No 1, 2020.

Sharma, S. and Jain, A., 'Role of sentiment analysis in social media security and analytics', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol 10, No 5, 2020.

Siddiqui, Z., 'Five Eyes intelligence chiefs warn on China's 'theft' of intellectual property', *Reuters*, 18 October 2023.

Siegmann, C. and Anderljung, M. 'The Brussels Effect and Artificial Intelligence: How EU Regulation Will Impact the Global AI Market', *Centre for the Governance of AI*, 16 August 2022.

Síthigh D. M. and Siems M., 'The Chinese social credit system: A model for other countries?', *The Modern Law Review*, Vol 82, No 6, 2019.

Soldatov, A., and Borogan, I., 'Russian Cyberwarfare: Unpacking the Kremlin's Capabilities', *Center for European Policy Analysis*, 8 September 2022.

Sprick, D., 'Predictive policing in China: An authoritarian dream of public security', *Naveiñ Reet: Nordic Journal of Law and Social Research (NNJLSR)*, No 9, 2019; A. Zenz and J. Leibold, 'Securitizing Xinjiang: police recruitment, informal policing and ethnic minority co-optation', *The China Quarterly*, Vol 242, 2020.

Stone, R., 'Iran's researchers increasingly isolated as government prepares to wall off internet', *Science*, 11 September 2023.

Strikwerda, L., 'Predictive policing: The risks associated with risk assessment', *The Police Journal*, Vol 94 No 3, 2020.

Su, Z., Cheshmehzangi, A., McDonnell, D.l, Bentley, B. L., Da Veiga, C. P. and Xiang, Y.T., 'Facial recognition law in China', *Journal of Medical Ethics*, Vol 48, No 12, 2022.

Sufi, F., 'Social Media Analytics on Russia–Ukraine Cyber War with Natural Language Processing: Perspectives and Challenges', *Information*, Vol 14, No 9, 2023.

Sun, R., Shi, L., Yin, C. and Wang, J., 'An improved method in deep packet inspection based on regular expression', *The Journal of Supercomputing*, Vol 75, 2019.

Sun, T. and Zhao, Q., 'Delegated censorship: The dynamic, layered, and multistage information control regime in China', *Politics & Society,* Vol 50, No 2, 2022.

Tabatabai, A. M., 'Other side of the Iranian coin: Iran's counterterrorism apparatus', *Journal of Strategic Studies*, Vol 41, No 1-2, 2018.

Taboada, M., 'Sentiment analysis: An overview from linguistics', *Annual Review of Linguistics*, Vol 2, 2016.

Tanczer, L. M., McConville, R., and Maynard, P., 'Censorship and surveillance in the digital age: The technological challenges for academics', *Journal of Global Security Studies*, Vol 1, No 4, 2016.

Thornton, R. and Miron, M., 'Towards the 'third revolution in military affairs' the Russian military's use of AI-enabled cyber warfare', *The RUSI Journal,* Vol 165, No 3, 2020.

Toh, M. and Erasmus, L., 'Alibaba's 'City Brain' is slashing congestion in its hometown', *CNN Business*, 15 January 2019.

Topor, L. and Tabachnik, A., 'Russian Cyber Information Warfare: International Distribution and Domestic Control', *Journal of Advanced Military Studies*, Vol 12, No 1, 2021.

Tschantret, J., 'Repression, opportunity, and innovation: The evolution of terrorism in Xinjiang, China', *Terrorism and political violence*, Vol 30, No 4, 2019.

Tucker, A., 'The citizen question: making identities visible via facial recognition software at the border', *IEEE Technology and Society Magazine*, Vol 39 No 4, 2022.

Tufekci, Z., 'Twitter and tear gas: The power and fragility of networked protest', Yale University Press (New Haven: United States of America), 2017.

Tulumello, S., and Lapaolo, F., 'Policing the future, disrupting urban policy today. Predictive policing, smart city, and urban policy in Memphis (TN)', *Urban Geography*, Vol 43, No 3, 2022.

UK government, 'The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023', Policy Paper, 1 November 2023.

UN, 'Programme of action to advance responsible State behaviour in the use of information and communications technologies in the context of international security: draft resolution / Albania, Argentina, Australia, Austria, Belgium, Bulgaria, Chile, Colombia, Croatia, Cyprus, Czechia, Denmark, Dominican Republic, Egypt, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Japan, Latvia, Lithuania, Luxembourg, Malta, Monaco, Netherlands, Norway, Paraguay, Poland, Portugal, Republic of Korea, Republic of Moldova, Romania, Senegal, Slovakia, Slovenia, Spain, Sweden, Switzerland, Tunisia, Türkiye, Ukraine, United Kingdom of Great Britain and Northern Ireland, United Republic of Tanzania and United States of America', UN Document, 2022.

UN Conference on Trade and Development, 'Science, Technology and Innovation Policy Review – Islamic Republic of Iran', UNCTAD, 2016.

UN Secretary-General's AI Advisory Body, 'Interim Report: Governing AI for Humanity', December 2023.

United Nations, Our Common Agenda: Report of the Secretary-General, 2021; United Nations, Our Common Agenda: Policy Brief 5. A Global Digital Compact — an Open, Free and Secure Digital Future for All, May 2023.

UNESCO, 'Recommendation on the Ethics of Artificial Intelligence', 16 May 2023.

United Nations, Universal Declaration of Human Rights, 1948.

Ünver, A. H., 'The Role of Technology: New Methods of Information, Manipulation and Disinformation, *Center for Economic and Foreign Policy Research*, 2023.

Ünver A. H. and Ertan, S. A., 'Politics of Artificial Intelligence Adoption Unpacking the Regime Type Debate', *Democratic Frontiers: Algorithms and Society*, M. Filimowicz (ed), Routledge (Oxfordshire: UK) 2022.

Ünver, A, H., 'The Logic of Secrecy: Digital Surveillance in Turkey and Russia', *Turkish Policy Quarterly*, Vol 17, No 2, 2018.

Ünver, A. H., 'Artificial intelligence, authoritarianism and the future of political systems', *Center for Economic and Foreign Policy Research*, 2018.

Ünver, A. H., Ertan, S. A., 'Democratization, state capacity and developmental correlates of international artificial intelligence trade', *Democratization*, 2023.

US Department of Security, DHS OIG Report: Secret Service and ICE Illegally Used Cell-Site Simulators, Electronic Privacy Information Center, 2023.

US government, 'China's data security law', International Trade Administration, 17 August 2021.

US National Science Foundation, 'Advancing Ethical Artificial Intelligence Through the Power of Convergent Research', nd.

Vainilavičius, J., 'Russia launches "Oculus" tool to monitor banned information online', *CyberNews*, 15 November 2023.

Varanasi, L., 'Putin says Russia will develop new AI technology to counter the Western monopoly, which he fears could lead to a 'digital abolition' of Russian culture', *Business Insider*, 26 November 2023.

Vasconcelos, H., Jörke M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., and Krishna, R., 'Explanations can reduce overreliance on AI systems during decision-making', *Proceedings of the ACM on Human-Computer Interaction*, Vol 7, No CSCW1, 2023.

Virki, T., 'Nokia Siemens to ramp down Iran operations', *Reuters*, 13 December 2011.

Votta, F., 'Algorithmic Microtargeting? Testing the Influence of the Meta Ad Delivery Algorithm', European Consortium for Political Research, Joint Sessions of Workshops, Sciences Po Toulouse, 25–28 April 2023.

Walker, S., 'Hungarian journalists targeted with Pegasus spyware to sue state', *The Guardian*, 28 January 2022.

Wang, T., Lu K., Chow, K. P., and Zhu, Q., 'COVID-19 sensing: negative sentiment analysis on social media in China via BERT model', Institute of Electrical and Electronics Engineers, Vol 8, 2020.

Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X., and Wang, F. Y., 'Generative adversarial networks: introduction and outlook', *IEEE/CAA Journal of Automatica Sinica*, Vol 4, No 4, 2017.

Wang, X., Jiang, W. and Luo, Z., 'Combination of convolutional and recurrent neural network for sentiment analysis of short texts', Technical papers, COLING 2016, the 26th international conference on computational linguistics, 2016.

Wang, X., Zhang, J., and Yan, W. Q., 'Gait recognition using multichannel convolution neural networks', *Neural computing and applications*, Vol 32, No 18, 2020.

Webster, G. and Laskai, L., 'Translation: Chinese Expert Group Offers 'Governance Principles' for 'Responsible AI'', DigiChina, Stanford University, 17 June 2019.

Wege, C. A., 'Iranian counterintelligence', *International Journal of Intelligence and Counterintelligence*, Vol 32, No 2, 2019.

West, S. M, 'Data capitalism: Redefining the logics of surveillance and privacy', *Business & Society*, Vol 58, No 1.

Wilde, G., 'Cyber Operations in Ukraine: Russia's Unmet Expectations', *Carnegie Endowment for International Peace*, 12 December 2022.

Wong, K. L. X. and Dobson, A. S., 'We're just data: Exploring China's social credit system in relation to digital platform ratings cultures in Westernised democracies', *Global Media and China*, Vol 4, No 2, 2019.

Workneh, T. W., 'Counter-terrorism in Ethiopia: manufacturing insecurity, monopolizing speech', *Internet Policy Review*, Vol 8, No 1, 2019.

Workneh, T. W., 'Digital cleansing? A look into state-sponsored policing of Ethiopian networked communities', *African Journalism Studies*, Vol 36, No 4, 2015.

World Economic Forum, 'Data Free Flow with Trust: Overcoming Barriers to Cross-Border Data Flows', White Paper, 16 January 2023.

World Economic Forum, 'Empowering AI Leadership: AI C-Suite Toolkit', 12 January 2022.

World Economic Forum, 'A Framework for Developing a National Artificial Intelligence Strategy', 4 October 2019.

World Economic Forum, 'Guidelines for AI Procurement', 2019.

Xu, X., 'To repress or to co-opt? Authoritarian control in the age of digital surveillance', *American Journal of Political Science*, Vol 65 No 2, 2021.

Xu, X., Kostka, and Cao, X., 'Information control and public support for social credit systems in China', *The Journal of Politics*, Vol 84, No 4, 2022.

Yalcintas, A. and Alizadeh, N., 'Digital protectionism and national planning in the age of the internet: the case of Iran', *Journal of Institutional Economics*, Vol 16, No 4, 2020.

Yang, E. and Roberts, M. E., 'The Authoritarian Data Problem', *Journal of Democracy*, Vol 34, No 4, 2023.

Yang, F., 'The tale of deep packet inspection in China: Mind the gap', in *2015 3rd International Conference on Information and Communication Technology,* IEEE, 2015.

Zeng, J., 'Artificial intelligence and China's authoritarian governance', *International Affairs*, Vol 96, No 6, 2020.

Zittrain, J. and Edelman, E. B., *'Internet filtering in China'*, *Institute of Electrical and Electronics Engineers Internet Computing*, Vol 7, No 2, 2003.

Zuboff, S. *The age of surveillance capitalism*, Routledge (London: United Kingdom), 2023.

8 Annexes

8.1 Techniques, tactics and procedures of algorithmic authoritarianism and bias: An overview of technical repertoires

It is important to outline AI-based repression techniques for the purposes of this IDA, as the technical repertoires that assist in repression are becoming increasingly complex. Algorithmic authoritarianism tools are becoming more diverse and less straightforward to detect as they leverage **dual-use (or general purpose) technologies** in the domains of AI, big data and sophisticated algorithms to monitor, manipulate and silence populations.

It is important to note that the most cutting-edge technologies are deployed by countries that have the technological prowess to use them at scale. This inevitably clusters the documented cases of use of these technologies around technologically more advanced countries (that tend to be more democratic – except for China), and those with a free press, which enables their dissemination. To that end, at the time of writing this IDA, authoritarian states have largely been deploying more 'old school' digital repression techniques such as vote injections, disinformation, censorship, or content manipulation and there are only a handful of documented cases of autocracies deploying advanced AI techniques for political purposes. This is largely because autocracies (again, except for China and to some extent Russia) lack the technological capacity to use them at scale; those who can deploy these techniques ultimately acquire their technologies and methods from China.

8.1.1 Automated Content Filtering (ACF)

ACF represents a collection of technologies and systems designed to analyse, categorise and potentially block or modify web content based on predefined parameters or algorithms. The primary intention behind these systems can range from the benign, such as protecting children from inappropriate content to more politically driven motives such as censoring oppositional views. However, the very same tools designed to protect users can also be deployed to filter out democratic expression, or views deemed as 'subversive'.

ACF often leverages advanced machine learning and NLP algorithms to understand the context and content of digital information[300]. This analysis can result in actions such as:

- Flagging content for human review
- Automatically blocking or restricting access
- Altering the visibility or rank of content in search results or social media feeds.

These algorithms can analyse text, images, videos and even audio, as many of the advanced content moderation algorithms used by social media platforms such as Meta, X or TikTok incorporate these multimedia ACF algorithms. Yet, these can also be deployed by authoritarian governments to weed out unwanted online discourse. Image recognition, for instance, can identify symbols or individuals that might be deemed undesirable by a governing authority. Similarly, text analysis can pinpoint specific terms, phrases, or sentiments that might be flagged as controversial or problematic.

In various authoritarian regimes, ACF plays a pivotal role in maintaining **political control and censorship.** By suppressing dissent, these tools can:

- Prevent citizens from accessing foreign news sources or oppositional local media.

[300] B. Mittelstadt, 'Automation, algorithms, and politics| auditing for transparency in content personalization systems', *International Journal of Communication*, No 10, 2016, p. 12; J. Herlocker, A. Konstan and J. Riedl, 'An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms', *Information retrieval*, Vol 5, 2002, pp. 287-310.

- Ensure a one-sided narrative, often bolstering the standing regime.
- Silence critics, activists and opposition figures by preventing their messages from reaching a broader audience[301].

ACF, while efficient for large-scale content moderation in many respects, can lead to overreach and a series of unintended consequences. By relying on algorithms to determine what content should be accessible or visible, various issues arise which can undermine the very objectives these systems were designed to achieve.

Algorithms, no matter how advanced, can make mistakes. When they detect patterns or keywords that match predefined 'offensive' or 'harmful' criteria, they may inadvertently block or remove legitimate, benign content. Such **false positives** can lead to the suppression of information that was never intended to be censored. Knowing that an automated system is constantly scanning and evaluating content, users might engage in self-censorship, avoiding certain topics or refraining from sharing genuine opinions for fear of repercussions[302]. This can **stifle public discourse** and the free exchange of ideas. If content creators are aware of the specific criteria that automated systems target, they might tailor their content to avoid these triggers, leading to a **homogenisation of content** where diversity of thought and expression is diminished[303].

Automated systems often struggle with **understanding context.** A word or phrase that is benign in one context might be flagged as offensive in another. Without human nuance, these systems can misinterpret content and lead to unjustified removals or blocks. For **businesses,** particularly online platforms, false positives can mean **lost revenues** if legitimate content becomes incorrectly flagged and removed. For **individuals,** being mistakenly targeted by these systems can lead to **social ostracisation** or in some contexts legal repercussions. Automated systems, especially those using complex machine learning models, can be **'black boxes',** making it hard to understand or challenge specific content filtering decisions. Without transparency, holding these systems accountable becomes challenging.

Any **biases** contained in data used to train these automated systems can be **perpetuated and amplified** by the systems themselves[304]. For instance, content from minority groups might be disproportionately flagged if any system's training data is skewed. There is also a risk of **over-dependence on technology.** Relying heavily on automated filtering might lead organisations or governments to believe that they have effectively addressed issues such as misinformation or hate speech when in reality they have merely treated some symptoms without addressing underlying causes[305]. There is a risk that entities with control over these systems might **manipulate them** for personal or political gain, by silencing opposition or promoting specific narratives. While ACF offers a scalable solution to manage vast amounts of data and content online, there are still certain flaws. Overreliance on such systems, without human oversight and regular audits, can lead to significant societal, cultural and economic implications.

---

[301] S. Jamil, 'Automated journalism and the freedom of media: Understanding legal and ethical implications in competitive authoritarian regime', *Journalism Practice*, Vol 17, No 6, 2023, pp. 1115-1138.

[302] A. Rauchfleisch and J. Kaiser, 'The false positive problem of automatic bot detection in social science research', PloS one, Vol 15, No 10, 2020.

[303] E. Nechushtai, R. Zamith, and S. C. Lewis, 'More of the Same? Homogenization in News Recommendations When Users Search on Google, YouTube, Facebook, and Twitter', *Mass Communication and Society*, 2023, pp. 1-27.

[304] E. Bozdag, 'Bias in algorithmic filtering and personalization. Ethics and information technology', Vol 15, 2013, pp.209-227.

[305] R. Binns, M. Veale, M. Van Kleek, and N. Shadbolt, 'Like trainer, like bot? Inheritance of bias in algorithmic content moderation', *Social Informatics: 9th International Conference, Proceedings*, Part II 9, 2017, pp. 405-415.

8.1.2 Sentiment analysis

Automated sentiment analysis, sometimes referred to as **opinion mining**[306], involves using natural language processing (NLP), text analysis and computational linguistics to ascertain the emotional tone or sentiment behind a series of words[307]. It is used to gain an understanding of public opinions, reactions and emotions regarding a specific topic, product, or event. The technology itself holds significant promise for numerous sectors, including business and marketing. However, its application by state and non-state actors has significant implications for citizens' freedom to access information.

8.1.2.1 Technical foundations of automated sentiment analysis

At its core, automated sentiment analysis is a specialised application of NLP, a domain of AI that seeks to enable machines to understand and interpret human language. By leveraging advanced machine learning techniques and vast datasets, sentiment analysis aims to decipher the underlying emotional tone of textual content[308]. The bedrock of effective sentiment analysis is **enormous, diverse and labelled textual data**. This data is often collected from sources such as social media platforms, customer reviews and fora. **Pre-processing** involves cleaning this data to remove noise (URLs or non-textual content, say), normalising text (for instance, converting everything to lowercase) and tokenisation (breaking sentences into individual words or phrases). In the context of sentiment analysis, **features,** such as specific words, phrases or even sentence structures can be recognised and extracted[309].

With the pre-processed data and features extracted, machine learning **models** are then **trained** to recognise patterns correlating with specific sentiments. Popular models for this purpose include **recurrent neural networks, long short-term memory networks** and language model transformers, the most popular of which is the Bidirectional Encoder Representations from Transformer **(BERT)**[310]. Modern NLP models are particularly adept at **understanding context**[311]. They can differentiate between multiple meanings of a word based on the surrounding text. This ability is crucial for accurate sentiment analysis, as many words can convey different sentiments based on their context.

Beyond basic positive or negative categorisations, advanced models can dissect text to **identify sentiments at various granularities.** This could mean distinguishing between different negative emotions (for example, sadness versus anger) or even gauging the intensity of a sentiment. Language is dynamic and evolves. As such, sentiment analysis systems must be **adaptable.** Continual or online learning allows these systems to refine their understanding based on new data, ensuring they remain accurate even as linguistic trends change. Sentiment analysis rarely operates in isolation. Its results are often **integrated into broader systems,** be it for business intelligence, customer relationship management, or even algorithmic content curation and this is why the use of such systems in political settings, such as political campaign messaging or censorship usually generates biased outcomes.

---

[306] H. Chen and D. Zimbra, 'AI and opinion mining', IEEE Intelligent Systems, Vol 25, No 3, 2010, pp.74-80.
[307] K. Ravi and V. and Ravi, 'A survey on opinion mining and sentiment analysis: tasks, approaches and applications', *Knowledge-based systems*, Vol 89, 2015, pp. 14-46.
[308] M. Taboada, 'Sentiment analysis: An overview from linguistics', *Annual Review of Linguistics*, Vol 2, 2016, pp. 325-347.
[309] R. Prabowo and M. Thelwall, 'Sentiment analysis: A combined approach', *Journal of Informetrics*, Vol 3, No 2, 2009, pp. 143-157.
[310] J. Devlin, M.W. Chang, K. Lee and K. Toutanova, 'Bert: Pre-training of deep bidirectional transformers for language understanding', ACL Anthology, 2018.
[311] X. Wang, W. Jiang and Z. Luo, 'Combination of convolutional and recurrent neural network for sentiment analysis of short texts', Technical papers, COLING 2016, the 26th international conference on computational linguistics, 2016, pp. 2428-2437.

8.1.2.2 Impacts on freedom to access information

Sentiment analysis, with its roots in market research and customer feedback systems, has evolved into a tool for analysing text to determine public sentiment. Its algorithms sift through online content – be it on social media, fora, or news comment sections – to gauge the emotional tone behind words. One of the significant advantages of sentiment analysis is its ability to **process huge amounts of data quickly.** This capability allows governments or organisations to gain insights into public opinion without having to conduct time-consuming surveys or face-to-face interviews. Sentiment analysis provides almost **instantaneous feedback.** For governments, especially in volatile political climates, this real-time pulse on public sentiment can offer crucial insights, allowing them to respond, adapt, or manipulate narratives promptly[312].

Advanced sentiment analysis tools do not just identify emotions; they can **also detect patterns and trends,** revealing how sentiment might change over time or in response to particular events. This temporal understanding can be instrumental for decision-makers in predicting potential public reactions. Sentiment analysis can be tailored to focus on specific demographic groups, regions, or even individual influencers. This **targeted approach** enables precise surveillance regarding dissent, especially in areas or among groups considered 'high risk'. However, in the hands of authoritarian regimes there are significant ramifications, in that sentiment analysis can be transformed from a tool of understanding to a **weapon of repression.** By quickly identifying pockets of dissent or unfavourable sentiment, governments can quickly implement repressive measures. Awareness of state-led sentiment monitoring can lead **individuals to self-censor,** fearing repercussions for expressing dissenting views. Over time, this could erode the free and open discourse fundamental to digital platforms.

8.1.3 Deep packet inspection

DPI is a sophisticated method of examining and managing network traffic. While traditional packet inspection looks only at the packet header (which contains meta information such as source and destination IP addresses), DPI scrutinises data packets' actual content as they pass an inspection point. This deeper look allows for more precise data filtering, surveillance and traffic management. However, when placed in the hands of authoritarian governments or misused by entities, DPI has profound implications for freedom of information and privacy.

DPI is a formidable tool in the arsenal of network management and surveillance. Unlike standard network monitoring tools which skim the surface, DPI provides deep granular visibility into data traffic, allowing a broader set of information to be extracted from digital communication and data transfers[313]. Traditional packet inspection systems, such as firewalls, generally examine only the header of a packet to determine its source, destination and protocol. DPI goes far beyond this, by **inspecting the data content itself.** Hence, it can parse details of the data being transmitted, be it an email, a webpage, or a voice call. One of the primary utilities of DPI is its ability to ascertain the specific **application** responsible for generating the traffic. By understanding the signature patterns of different applications, DPI can distinguish between a video stream from a site such as YouTube and a voice call made via Skype, even if they use the same basic protocols. This granularity facilitates targeted network management and in some contexts surveillance.

DPI is not just about identifying the type of application; it can be tuned to look for specific **keywords or patterns** within transmitted data. For example, an authoritarian regime might configure DPI tools to flag communications containing politically sensitive terms or phrases[314]. This content-based filtering allows for

---

[312] S. Sharma and A. Jain, 'Role of sentiment analysis in social media security and analytics', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol 10, No 5, 2020.

[313] R. T.El-Maghraby, N. Abd Elazim and A. Bahaa-Eldin, 'A survey on deep packet inspection', 12th International Conference on Computer Engineering and Systems, 2017, pp. 188-197.

[314] C. Fuchs, 'Societal and ideological impacts of deep packet inspection internet surveillance', *Information, Communication & Society*, Vol 16, No 8, 2013, pp. 1328-1359.

both real-time surveillance and subsequent action, such as data collection, blocking, or user flagging. Beyond surveillance, DPI has practical applications in network management. By recognising the type and content of data packets, DPI can be used to prioritise or throttle specific kinds of **traffic.**

8.1.4 Facial recognition and surveillance

Facial recognition is a subset of biometric technology that identifies or verifies an individual based on patterns in facial features. With the advent of deep learning and expansion in the computational power of modern systems, facial recognition technologies have achieved significant advancements in accuracy and speed[315]. However, their applications in surveillance have raised serious human rights and privacy concerns, as they provide governments and organisations with unprecedented power to track and monitor individuals.

8.1.4.1 Technical foundations of facial recognition in surveillance

Facial recognition has emerged as a cutting-edge tool in the arsenal of surveillance technologies. Its rise can be attributed to both advancements in camera technology and breakthroughs in AI. The initial step in any facial recognition system is **the acquisition of visual data.** High-resolution cameras, often equipped with infrared capabilities to capture images in low-light conditions, are strategically placed in public spaces, entry points and other areas of interest. These cameras feed images continuously, either in real-time for immediate analysis or as stored data for subsequent processing[316].

Once raw images are available, the next challenge is identifying faces within them, which is especially difficult in crowded or dynamic environments. However, by zoning in on areas that resemble **human faces,** advanced algorithms are so sophisticated that they can **detect** different outlines in a single frame, even if partially obscured or at different angles[317]. Once a face has been detected, the system then delves into the finer details to **extract the face 'features'.** This step is crucial, for it is here that unique facial landmarks are identified and quantified. These can include: the contours of the eye sockets; the width of the nose bridge; the depth of the cheekbones; and the curvature of the lips. Such measurements, often numbering in the hundreds, constitute a face's unique 'signature'[318].

With this signature, the system then embarks on the **task of matching**[319]. A database of facial signatures (often tied to identities) serves as the reference. The system compares the newly extracted signature against this database, seeking a match. The evolution of **deep learning,** a subset of AI, has been a game-changer for facial recognition. Traditional algorithms often faltered with variations in lighting, angles, or facial expressions. However, **convolutional neural networks** – a type of deep learning model – trained on millions of facial images, have brought about significant improvements in accuracy.[320] They can detect and recognise faces with a variety of expressions, head positions and lighting conditions. Their ability to learn

[315] M. Gray, 'Urban surveillance and panopticism: will we recognize the facial recognition society?', *Surveillance & Society*, Vol 1, No 3, 2003, pp. 314-330.
[316] M. Gray, 'Urban surveillance and panopticism: will we recognize the facial recognition society?', *Surveillance & Society*, Vol 1, No 3, 2003, pp. 314-330.
[317] V. D. A. Kumar, S. Malathi, K. Vengatesan and M. Ramakrishnan, 'Facial recognition system for suspect identification using a surveillance camera', *Pattern Recognition and Image Analysis*, Vol 28, 2018, pp. 410-420.
[318] L. Introna and D. Wood, 'Picturing algorithmic surveillance: The politics of facial recognition systems', *Surveillance & Society*, Vol 2, No 2/3, 2004, pp. 177-198.
[319] A. M. Burton, S. Wilson, M Cowan, V. and Bruce, 'Face recognition in poor-quality video: Evidence from security surveillance. Psychological Science', Vol 10, No 3, 2018, pp. 243-248.
[320] S. Almabdy and L. Elrefaei, 'Deep convolutional neural network-based approaches for face recognition', *Applied Sciences*, Vol 9, No 20, 2019, pp. 4397.

from vast amounts of data means they continually refine their accuracy, distinguishing even subtle differences between faces.

8.1.4.2 Applications and impacts on repression

Facial recognition technology, driven by advancements in AI, has been rapidly adopted by various governments around the world to conduct mass surveillance and monitoring. While the technology holds promise for *inter alia* enhancing public safety and streamlining administrative functions, its deployment in surveillance networks naturally raises concerns about privacy, civil liberties and potential misuse[321]:

- As discussed in greater detail within the case studies, China stands out for its ambitious adoption of facial recognition in surveillance. The **'Skynet'** initiative is a testament to this, aiming at pervasive video surveillance coverage in urban centres[322]. In regions such as Xinjiang, this technology takes on a more ominous tone, being used as a tool for stringently monitoring and controlling the Uighur Muslim population. Such practices have drawn international criticism and concerns over gross human rights violations.
- Moscow has rolled out one of the most extensive camera systems in Europe. Bolstered by facial recognition capabilities, these cameras aid in crime detection and public safety measures. However, they have also been used to monitor and sometimes detain opposition figures and participants in public protests. The blending of facial recognition with surveillance networks here has ignited debates on personal freedoms in an already politically charged environment.
- India has ventured into using facial recognition for a variety of administrative and security purposes. From the push for Aadhaar, a biometric identification system, to initiatives in policing, the technology is gaining ground. However, in the absence of robust data protection laws, there are increasing concerns about how this biometric data might be used or misused, especially in tracking protesters and dissenters. In certain cases, the Indian government was accused of using Aadhaar to target political dissidents deliberately using facial biometric data[323].
- Various American cities, including San Francisco and Boston, have banned the use of facial recognition by local government and police, citing potential misuse and biases inherent in the technology. However, at the federal level, agencies such as the FBI maintain vast facial recognition databases. The technology's deployment for border inspections as well as airport identification and security checks are also expanding, with companies and federal agencies exploring its potential. ICE has faced criticism for policies that increase the risk of mistaken detainment, deportation, racial profiling and discrimination against immigrant communities using facial recognition technology. Notably, since 2002, ICE has mistakenly identified at least 2 840 U.S. citizens for deportation, with some estimates suggesting over 20 000 such cases between 2003 and 2010. These wrongful detentions often occur without proper legal representation or due process. Examples include the cases of Davino Watson, a U.S. citizen detained for over three years, and Peter Sean Brown, detained for three weeks due to mistaken identity resulting from poorly calibrated facial recognition systems. The absence of retrospective justice and correction of

[321] G. Kostka, L. Steinacker and M. Meckel, 'Under big brother's watchful eye: Cross-country attitudes toward facial recognition technology', *Government Information Quarterly*, Vol 40, No 8, 2023.
[322] J. Leibold, 'Surveillance in China's Xinjiang region: Ethnic sorting, coercion, and inducement', *Journal of contemporary China*, Vol 29, No 121, 2020, pp. 46-60.
[323] P. Dixon, 'A Failure to "Do No Harm" -- India's Aadhaar biometric ID program and its inability to protect privacy in relation to measures in Europe and the U.S.', *Health Technol (Berl)*, National Institutes of Health, 2017.

wrongdoings highlight significant concerns regarding ICE's enforcement practices and their impact on minority communities[324].

- Brazil has employed facial recognition technology during large events, such as the Rio Carnival and football matches, to identify criminals or individuals with outstanding warrants. As the technology spreads, its use in everyday surveillance is growing, especially in crime-intensive regions. However, despite proclaimed benefits in crime prevention, concerns persist about its accuracy and implications for innocent civilians being misidentified, with a recent case about the Brazilian police using AI-based facial recognition tools without proper governmental authorisation to target dissenting groups[325].
- With Chinese assistance, Zimbabwe has begun integrating facial recognition technology into its urban surveillance infrastructure. Opposition figures and human rights activists have expressed concerns that the government is actively using this technology to monitor and suppress opposition activities, especially during election periods or public demonstrations[326].
- In 2019, reports emerged that Ecuador was considering implementing a system similar to China's 'Skynet'. The idea was to integrate 4 300 cameras with facial recognition technology throughout the country. While the primary goal was safety and security, concerns were raised about the potential for its use in naming and shaming individuals for various infractions, from traffic violations to more personal misdemeanours[327].
- South Africa has seen an increasing number of security cameras equipped with facial recognition in public areas, primarily for crime prevention. However, there is potential for misuse, as these systems can easily be repurposed to display the faces of individuals engaged in minor infractions, using public shame as a deterrent[328].

The rise of these technologies worldwide reflects a broader move towards technologically enforced social compliance. It also underscores the importance of defining boundaries and regulations for the use of such technologies, ensuring that they do not infringe on personal rights or contribute to unwarranted public humiliation.

8.1.5 Predictive policing

By harnessing the power of data analysis and machine learning algorithms, predictive policing seeks to prognosticate potential crime hotspots, potential perpetrators and even likely victims. Advocates of this technology proclaim its capabilities as a revolutionary advancement, enabling more efficient allocation of police resources and a proactive approach to crime prevention[329]. However, this predictive mechanism might perpetuate, or even exacerbate, pre-existing biases found within historical crime data. This can lead to a self-fulfilling prophecy where certain communities face heightened surveillance and scrutiny, thereby further embedding patterns of repression and discrimination[330]. In what follows, this section aims to

[324] D. Mehrotra, 'ICE Records Reveal How Agents Abuse Access to Secret Data', *Wired*, 17 April 2023.
[325] Associated Press, 'Brazil police conduct searches targeting intelligence agency's use of tracking software', 20 October 2023.
[326] F. S. Matiashe, 'Zimbabwe's cyber city: Urban utopia or surveillance menace?', *Reuters*, 21 February 2023.
[327] C. Rollet, 'Ecuador's All-Seeing Eye Is Made in China', *Foreign Policy*, 9 August 2018.
[328] K. Hao and H. Swart, 'South Africa's private surveillance machine is fueling a digital apartheid', MIT Technology Review, 19 April 2022.
[329] A. Meijer and M. Wessels, 'Predictive policing: Review of benefits and drawbacks', *International Journal of Public Administration*, Vol 42, No 12, 2019, pp. 1031-1039.
[330] M. Kaufmann, S. Egbert and M. Leese, 'Predictive policing and the politics of patterns', *The British journal of criminology*, Vol 59, No 3, 2019, pp. 674-692.

provide a comprehensive examination of these dual perspectives, highlighting the global impacts and the ethical quandaries posed by such technology.

8.1.5.1 Technical foundations of predictive policing

The rise of predictive policing is rooted in its ability to harness vast volumes of data, transforming them into actionable insights for law enforcement agencies. Drawing from a plethora of sources such as reported crimes, demographic patterns, online chatter on social media and even atmospheric variations, predictive policing attempts to foretell where and when crimes might occur, as detailed here.

This foundational step **aggregates vast datasets** from disparate sources[331]. For instance, the Los Angeles Police Department's 'Operation **LASER'** programme leveraged data from gunshot detection systems, crime reports and gang territory maps, but was ultimately shut down following public opposition to its discriminatory and biased algorithms[332]. Similarly, the UK's Kent Police used social media activity as an input for its predictive model, alongside traditional crime data and was shut down soon after as the details of its data processing methods became public knowledge[333].

Before being ingested by algorithms, data often requires **substantial refinement.** It must be cleaned to remove outliers or irrelevant information and might be standardised or transformed to ensure compatibility. For instance, the Chicago Police Department's 'Strategic Subject List' processed arrest records, victim data and other crime-related issues to rank individuals based on their likelihood of being involved in violent crimes. This was withdrawn following the exposure of biases in its prediction system by civil liberties NGOs[334].

At this stage, modern statistical methods or advanced machine learning algorithms sift through the data, looking for patterns. Techniques such as regression analysis, clustering, or neural networks might be used for **predictive modelling.** In Memphis, the **'Blue CRUSH'** (Crime Reduction Using Statistical History) initiative employed IBM's Statistical Package for the Social Sciences Modeler software to identify patterns and make predictions about future crime hotspots. Blue CRUSH was withdrawn and disbanded after a lengthy legal process between the victims and the state police authority, citing systematic bias and abuse[335].

While the information above gives an overview of the technical flow, it is imperative to understand that predictive policing's efficacy and ethical standing are topics of ongoing debate, especially given concerns about data privacy, inherent biases in data and the potential for reinforcing prejudiced policing practices.

8.1.5.2 Applications and repression implications

Reinforcement of existing biases

Machine learning and predictive analytics, while revolutionary in many domains, are beholden to the data on which they are trained. As mentioned earlier, historical biases embedded within this data can result in these tools perpetuating and sometimes exacerbating discriminatory practices.

---

[331] W. Hardyns and A. Rummens, 'Predictive policing as a new tool for law enforcement? Recent developments and challenges', *European journal on criminal policy and research*, Vol 24, 2018, pp. 201-218.
[332] J. Bhuiyan, 'LAPD ended predictive policing programs amid public outcry. A new effort shares many of their flaws', *The Guardian*, 8 November 2021.
[333] L. Strikwerda, 'Predictive policing: The risks associated with risk assessment', *The Police Journal*, Vol 94 No 3, 2020, pp. 422-436.
[334] J. Saunders, P. Hunt and J. S. Hollywood, 'Predictions put into practice: a quasi-experimental evaluation of Chicago's predictive policing pilot', *Journal of experimental criminology,* Vol 12, 2016, pp. 347-371.
[335] S. Tulumello and F. Lapaolo, 'Policing the future, disrupting urban policy today. Predictive policing, smart city, and urban policy in Memphis (TN)', *Urban Geography*, Vol 43, No 3, 2022, pp. 448-469.

- In cities across the USA where predictive policing has been implemented, there have been instances where the tools have disproportionately targeted minority and marginalised communities. For instance, a system deployed in Oakland suggested police should allocate more resources to regions already heavily policed for drug offences, neighbourhoods predominantly inhabited by racial minorities[336].
- In China, the government employs predictive policing in Xinjiang, using a system called the **Integrated Joint Operations Platform.** This system gathers extensive data on citizens, including their biometrics, online activities and location data, to identify and monitor individuals deemed as potential threats. The biometric and credit score data collected as part of the Integrated Joint Operations is then used to conduct statistical analyses on the likelihood of criminal activity, rendering entire ethnic and religious groups stigmatised and disproportionately represented by the algorithm[337].
- In 2017, the German state of Bavaria introduced a predictive policing software named **'PREDPOL'.** Concerns were raised about its potential to magnify historical biases since the system was primarily fed data on reported crimes, potentially sidelining unreported or differently perceived offences[338].

Suppression of dissent and political activities

Predictive tools can be appropriated by authoritarian regimes to foresee political uprisings, protests, or even trends that do not align with the state's ideology, leading to pre-emptive actions against would-be dissidents:

- Recent protests in China, largely driven by frustration with the strict 'zero-COVID' policy, have met with a significant police response. In cities such as Beijing and Shanghai, there has been a notable increase in police presence on the streets, aimed at preventing further protests. Universities have also been sending students home as part of efforts to tighten COVID restrictions and discourage gatherings that could lead to dissent. Human rights groups have reported a ramping up of 'collective punishment' against activists and dissidents, impacting not only the individuals involved but also their families. This has included exit bans, detentions, evictions and other forms of harassment, which are seen as part of a broader crackdown on dissent both within China and beyond its borders. All these monitoring practices were fuelled by facial recognition and smart policing applications[339].
- The 'Fatherland Card' **(Carnet de la Patria)** in Venezuela has indeed evolved into a tool that extends beyond its initial purpose of facilitating access to public services. The card, which incorporates a unique personalised QR code and functions as a digital wallet, was initially introduced to streamline the distribution of food and other state-administered services. However, over time it has become deeply integrated into various state processes, including access to legal and personal documents, which are notoriously difficult to obtain in Venezuela. More than 70 % of Venezuelans reportedly carry the Fatherland Card, including both supporters and opponents of

[336] P. J. Brantingham, M. Valasik, and G. O. Mohler 'Does predictive policing lead to biased arrests? Results from a randomized controlled trial', *Statistics and public policy*, Vol 5, No 1, 2018, pp. 1-6.
[337] D. Sprick, 'Predictive policing in China: An authoritarian dream of public security', *Naveiñ Reet: Nordic Journal of Law and Social Research (NNJLSR)*, No 9, 2019; A. Zenz and J. Leibold, 'Securitizing Xinjiang: police recruitment, informal policing and ethnic minority co-optation', *The China Quarterly*, Vol 242, 2020, pp. 324-348.
[338] S. Egbert, 'About discursive storylines and techno-fixes: the political framing of the implementation of predictive policing in Germany', *European Journal for Security Research*, Vol 3, 2018, pp. 95-114.
[339] F. Jiang and C. Xie, 'Roles of Chinese police amidst the COVID-19 pandemic', *Policing: A Journal of Policy and Practice*, Vol 14, No4, pp. 1127-1137.

the current political regime[340]. The card's benefits have grown to include access to government bonds and fuel discounts; furthermore, it is now required for accessing housing bonds and pension payments.

This expansion of uses has raised concerns about the card's potential role in social control and citizen monitoring. Experts and human rights groups have expressed concerns that the Fatherland Card could be used for monitoring and controlling the population using algorithmic practices such as statistical inferences about human behaviour and profiling. These concerns are grounded in the card's capabilities to store and transmit extensive personal data about its holders to government servers. The database associated with the card system reportedly includes details such as: birthdays; family information; employment and income; property owned; medical history; state benefits received; social media presence; political party membership; and voting activity. The involvement of Chinese telecom giant ZTE has been a crucial element in the process of developing the fatherland database and creating a mobile payment system for use with the card[341]. During elections, the presence of 'Fatherland Card booths' near polling stations has been reported, where cardholders were encouraged to register and promised access to food and subsidy bonuses. This has led to allegations of the card being used to influence voter behaviour and discriminate against those without it.

8.1.6 Deepfake technology

Deepfake technologies utilise advanced AI, specifically deep learning, to create hyper-realistic, but entirely fake content. By mimicking the appearance and voice of individuals, these tools can produce videos that appear to show them saying or doing things that they never did. As these technologies become more sophisticated and accessible, they pose unique threats to elections and human rights worldwide.

8.1.6.1 Technical foundations of deepfakes

Deepfakes, a confluence of 'deep learning' and 'fake', represent one of the most advanced and concerning evolutions in synthetic media. At the foundational level, deepfakes are generated using a structure known as **Generative Adversarial Networks (GANs),** the technical intricacies of which are detailed below.

GANs consist of two neural networks, the generator and the discriminator, which are pitted against each other in a kind of game[342].

- Generator: This network takes in random noise as an input and produces data (images).
- Discriminator: This network evaluates the data, attempting to distinguish between genuine and synthetic.

During the training phase:

- The generator creates a new piece of data.
- The discriminator evaluates it against genuine data.
- Based on the discriminator's assessment, the generator adjusts its parameters to produce more convincing data in the next iteration.

---

[340] J. Ragas, 'A starving revolution: ID cards and food rationing in Bolivarian Venezuela', *Surveillance & Society*, Vol 15, Nos 3 and 4, 2017, pp. 590-595.

[341] ABC/Reuters, 'Chinese telecom giant ZTE 'helped Venezuela develop social credit system'', *ABC News*, 16 November 2018.

[342] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta and A. A. Bharath, 'Generative adversarial networks: An overview', *IEEE signal processing magazine*, Vol 35, No 1, 2018, pp. 53-65.

- This process is repeated, often for thousands or even millions of iterations, until the generator produces data that the discriminator cannot distinguish from real data.

Deepfakes necessitate **substantial computational power** and vast **datasets.** For instance, to create a convincing deepfake video of a person, training data would ideally include numerous images or videos of that person from various angles, lighting conditions and facial expressions. Modern Graphic Processing Units or even specialised tensor processing units are often employed to expedite the computationally intensive training process[343].

Over time, various architectures and **refinements** have been proposed to enhance the quality and reduce the training time of GANs. Some of these include Conditional GANs, **CycleGANs** and Progressive Growing GANs[344]. Each comes with its own set of advantages, depending on the specific application or desired outcome. While the first deepfakes primarily focused on manipulating **video footage,** advances in **audio** synthesis have also been made. Tools such as *DeepVoice* or *WaveNet* can mimic human voices and when paired with deepfake video, can create a highly realistic audio-visual fake. Given their technical sophistication and the increasing ease with which they can be created, deepfakes present substantial challenges in areas such as misinformation, digital forensics and security.

8.1.6.2 Applications and repression implications

The dawn of deepfakes has created an unprecedented **challenge** in discerning fact from fiction, especially in the sensitive domain of politics. With **elections** representing the pinnacle of democratic processes, potential manipulation through deepfakes is a cause for global concern, as the following example highlights.

One of the more insidious uses of deepfakes is the ability to **cast doubt on real, authentic footage** or information. This approach leverages the very existence of deepfake technology as a defence mechanism. In 2019, when President Ali Bongo of Gabon appeared seemingly 'different' in a New Year's address, sceptics were quick to label the video as deepfake[345]. This narrative, whether accurate or not, triggered political unrest, leading to an attempted coup, as opponents cited concerns about a potential political vacuum.

The ability to **forge realistic videos** can be **weaponised to depict individuals in compromising** situations falsely, thereby undermining their credibility and potentially altering the course of political or public sentiment. In a scandal that rocked Malaysia in 2019, videos purporting to show a cabinet minister engaged in intimate acts with another man emerged. Amidst the controversy which followed, some observers and experts speculated the potential use of deepfakes aimed at sabotaging the minister politically.

Deepfakes can **alter public statements** of politicians or influencers, changing the course of political campaigns or **affecting public sentiment** and even though they may not affect elections directly, they may influence voter sentiment and voter behaviour. The Wagner Group rebellion in Russia in 2023, led by

---

[343] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F. Y. Wang, 'Generative adversarial networks: introduction and outlook', *IEEE/CAA Journal of Automatica Sinica*, Vol 4, No 4, 2017, pp. 5 88-598.

[344] Conditional GAN, commonly abbreviated refers to a variant of GAN (Generative Adversarial Network) that specialises in generating images conditionally through a generator model. The Cycle Generative Adversarial Network, often referred to as CycleGAN, represents a method used to train deep convolutional neural networks specifically for the purpose of translating between different types of images. Progressive Growing GAN enhances the GAN training method by starting with small images and progressively adding layers to the generator and discriminator models, thereby enabling stable training and the production of large, high-quality images until the target size is reached. Please see: T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, 'Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion', *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6820-6824.

[345] S. Cahlan, 'How misinformation helped spark an attempted coup in Gabon', *The Washington Post*, 13 February 2020.

Yevgeny Prigozhin, marked a significant conflict within the country. This event was characterised by Prigozhin's forces taking control of key locations and advancing towards Moscow. The rebellion, which was a response to tensions between the Wagner Group and the Russian Ministry of Defence, was eventually settled through an agreement brokered by Belarusian President Alexander Lukashenko. The use of deepfakes in such a context, showing manipulated images or videos of figures like President Vladimir Putin, had significant implications and because of the resulting confusion prevented the mobilisation of certain Russian regiments[346]. Deepfakes in this scenario were used to spread misinformation, create confusion, or manipulate public opinion against Putin, exacerbating the emergency. When used in an election or an emergency setting, deepfakes can significantly alter public sentiment and sow mistrust towards governments[347].

Perhaps the biggest cautionary tale on the potential damage of deepfakes on democratic elections happened in Slovakia. Just before Slovakia's elections on 30 September 2023, a deepfake audio recording on Facebook purportedly featured Michal Šimečka of the Progressive Slovakia party discussing election rigging. Fact-checkers indicated AI manipulation, but Slovakia's election rules hindered widespread debunking. This occurred during a critical election between pro-NATO and anti-NATO parties, highlighting AI's potential to disrupt elections. Despite Meta's efforts to label and down-rank such content, this incident underscores the challenges fact-checkers face against AI-manipulated media, especially with limited tools to detect such manipulations effectively. This example serves as a caution for countries facing future elections[348].

The proliferation of deepfakes underscores the need for advanced verification tools and heightened public awareness. As the line between genuine content and fabricated media continues to blur, the need for digital literacy, forensic tools and stringent legislative measures becomes increasingly vital.

8.1.7 Gait detection

Gait detection, also known as gait recognition or gait analysis, is a biometric method that identifies individuals based on the way they walk[349]. This technology has become increasingly sophisticated and is used in various fields, including security and surveillance. Unlike other biometric technologies such as facial recognition or fingerprint identification, gait detection does not require direct contact with the subject and can be effective even at a distance or in low visibility conditions. This method is built on the assumption that each person has a distinct way of walking and moving limbs, resulting in a temporal analysis of these movements being automatically able to identify the person being tracked[350].

The technical foundations of gait detection technology form a complex and multifaceted process, beginning with the acquisition of data and culminating in the matching of gait patterns based on pre-set parameters. The transition from raw data to identifiable gait patterns involves various critical stages, each contributing to the overall efficacy of this technology:

- The initial stage in gait detection is data acquisition. This crucial step typically involves the use of advanced surveillance cameras, such as CCTV and 3D models, strategically placed to capture the walking patterns of individuals. These cameras are designed to record the intricate details

[346] V. Bahl, 'No, this video doesn't show the Wagner Group moving to Belarus', *France24*, 11 July 2023; C. Marchant de Abreu 'These images don't show confrontations between Wagner group and Russian army', *France24*, 26 June 2023.
[347] J. Reid, 'Putin quizzed by apparent AI version of himself during live phone-in', *Meta CNBC*, 14 December 2023.
[348] M. Meaker, 'Slovakia's Election Deepfakes Show AI Is a Danger to Democracy', *Wired*, 3 October 2023.
[349] T. K. Lee, M. Belkhatir, and S. Sanei, 'A comprehensive review of past and present vision-based techniques for gait recognition', *Multimedia tools and applications,* Vol 72, 2014, pp. 2833-2869.
[350] I. Bouchrika, 'A survey of using biometrics for smart visual surveillance: Gait recognition. Surveillance in Action,' *Technologies for Civilian Military and Cyber Surveillance*, 2018, pp. 3-23.

of a person's gait, even in varying environmental conditions and from different angles, providing a comprehensive dataset for further analysis[351].

- Once this gait data is captured, it undergoes a rigorous process of signal processing and pattern recognition. In this phase, the raw data from the video footage is processed to identify and extract meaningful patterns. This process includes breaking down the gait into various segments such as stride length, speed and rhythm. To achieve this, signal processing techniques are employed, including time-series analysis and frequency domain analysis. These methods allow for a detailed breakdown of the gait cycle, highlighting unique features and patterns that can be used for identification purposes[352].
- The core of gait detection technology lies in its use of machine learning algorithms. These algorithms are the driving force behind the identification and classification of different gait patterns. They are trained using a vast dataset of known gait patterns, learning the nuances and variations that distinguish one individual's walk from another. Smart city bundles that contain extensive CCTV camera networks are ideal data collection mechanisms for this purpose, as they can capture thousands of citizens daily, through various angles and times of the day. Once trained, these algorithms apply their learned knowledge to new samples of gait data, classifying and identifying them with increasing accuracy. Popular techniques in this domain include **neural networks, support vector machines** and **deep learning,** each contributing to the system's ability to discern and classify gait patterns effectively[353].
- The final step in this technological process is **feature extraction** and **matching.** Here, key features of an individual's gait, such as limb movement, body posture and dynamic weight distribution, are carefully extracted from the processed data. These features are critical in defining the uniqueness of each individual's gait. The extracted features are then compared against a pre-existing database, searching for matches. This comparison is a delicate task, requiring high precision to ensure accurate identification. The matching process is not just about finding identical matches, but also about recognising patterns even in the presence of minor variations or environmental changes.

8.1.7.1 Applications and repression implications

In the realm of surveillance and political control, the application of gait detection technology has become a topic of both significant interest and concern. Its use spans various domains, from enhancing mass surveillance capabilities to targeting specific groups, such as political dissidents. This technology's integration into broader security systems, alongside its implications for privacy and ethical considerations, raised alarms in the landscape of surveillance and personal freedom.

Gait detection is rapidly becoming a cornerstone in mass surveillance, especially in densely populated public areas such as airports, railway stations and urban centres. China is currently the only large-scale user of this technology. It has been using gait detection in major cities and around key transport hubs such as railway stations and airports for crime prevention purposes[354], although even before its deployment in 2017 the technology was largely trained and experimented with in Xinjiang[355]. The ability of this

[351] S. Maity, Abdel-Mottaleb M. and S. S. Asfour, 'Multimodal low resolution face and frontal gait recognition from surveillance video', *Electronics*, Vol 10, No 9, 2021, p. 1013.
[352] Y. Makihara, M. S. Nixon, and Y. Yagi, 'Gait recognition: Databases, representations, and applications', *Computer Vision: A Reference Guide*, 2020, pp. 1-13.
[353] X. Wang, J. Zhang, and W. Q. Yan, 'Gait recognition using multichannel convolution neural networks', *Neural computing and applications*, Vol 32, No 18, 2020, pp. 14275-14285.
[354] D. Kang, 'Chinese 'gait recognition' tech IDs people by how they walk', *Associated Press*, 6 November 2018.
[355] P. Mozur, 'Inside China's Dystopian Dreams: A.I., Shame and Lots of Cameras', *The New York Times*, 8 July 2018.

technology to track and monitor individuals unobtrusively, often without their consent or awareness, presents a powerful tool for authorities. This capability allows for the constant monitoring of people's movements and behaviours, potentially altering the dynamics of public spaces and personal privacy.

One of the more controversial applications of gait detection is in the identification and tracking of political dissidents. In certain countries, authorities have employed this technology to scrutinise footage from protests or public gatherings. By doing so, they can identify individuals based on their unique gait, enabling them to target and track political opponents discreetly. This application raises profound concerns about the suppression of dissent and freedom of expression.

Border security is another area where gait recognition is increasingly being deployed. It serves as an additional layer of security, identifying individuals who might be on watchlists, even if they attempt to alter or disguise their facial features. Russian Ministry of Internal Affairs is currently experimenting with border monitoring applications of gait detection[356]. Gait detection is often integrated with other biometric technologies, such as facial recognition, to create a more comprehensive surveillance package. This integration offers a multi-faceted approach to identification and tracking, combining different biometric markers for more accurate and robust surveillance capabilities.

However, this amalgamation also intensifies the potential for invasive monitoring and raises the stakes for privacy invasion[357]. The use of gait detection in political settings has sparked significant ethical and privacy concerns. The technology's potential for indiscriminate mass surveillance and misuse against political adversaries poses critical questions about the balance between security and individual rights. These concerns are further compounded by the lack of transparency and consent in the deployment of such surveillance technologies[358]. Moreover, the deployment of gait detection technologies in political contexts is increasingly being scrutinised under legal and regulatory frameworks. These frameworks vary widely across jurisdictions, with some areas calling for stricter regulations to prevent abuse and ensure the responsible use of such technologies.

8.2 Current trends in AI abuse for repression

8.2.1 Outcomes and motivations: Why do governments engage in algorithmic authoritarianism?

Governments' engagement in algorithmic and AI-based repression practices is deeply rooted in their desire for control, stability and power. In many cases, the allure of these technologies is their promise of efficiency and precision not only in monitoring but also in suppressing dissenting voices, especially in digital spaces. Historically, maintaining control over narratives, information flow and public sentiment have been the hallmarks of authoritarian rule. With the digital age, the terrain has simply shifted from physical spaces and traditional media to online fora and social networks.

AI and algorithms offer governments unprecedented capabilities to sift through vast amounts of data quickly. Most authoritarian governments have shifted their focus away from merely collecting big data', defined as the volume of information not processable by standard computer hardware or storage infra-structures, to fathoming and interpreting what is contained therein. For an authoritarian regime, this means identifying patterns of dissent, potential threats, or even understanding public sentiment to a

[356] L. Pascu, 'Russian Ministry testing gait recognition as part of national biometric surveillance system', BiometricUpdate.com, 25 February 2020.
[357] European Parliament, Biometric Recognition and Behavioural Detection, Briefing for the JURI and PETI committees, PE 697.131, September 2021.
[358] European Data Protection Board, Guidelines 05/2022 on the use of facial recognition technology in the area of law enforcement, Version 1.0, 2022.

granularity that was previously unimaginable[359]. This necessitates deploying AI not just for data collection purposes, including more advanced biometric and gait-based information, but also for calculation, prediction and large-scale statistical analysis. **Large language models (LLMs), natural language processing (NLP)** and **computer vision,** which have so far remained at the frontiers of scientific research, are now at the core of governments' interests and planning for high-technology research and development as well as imports[360]. When political activists, journalists, or regular citizens use digital platforms to voice opposition, share news, or organise movements, their digital footprints become sources of invaluable data for governments aiming to suppress such actions. To that end, live streaming of social media, the internet of things, GPS trackers and wearable biometric devices are gradually turning into government surveillance sensors.

Moreover, control over information flow has always been a powerful tool in the arsenal of any regime. By employing AI-driven content filters, governments can ensure that only state-approved narratives gain prominence, effectively drowning out or outright blocking dissenting views. These tools not only help in the active suppression of undesirable content but can also be used to propagate state-sanctioned content, leading to a more passive form of manipulation where citizens are fed a curated version of reality.

The cover of law enforcement and counterterrorism often provides a convenient justification for the expansion of these repressive technologies. Many governments assert that such tools are essential for national security, to fight against external threats, or to maintain internal stability. While there might be genuine cases where AI can aid legitimate law enforcement efforts, the line between genuine security concerns and political repression often becomes blurred, leading to an environment where the technology reinforces authoritarian tendencies under the guise of maintaining order. This also blurs the line between authoritarian and democratic governments' deployment of AI-based surveillance systems, as both regime types often use similar justifications or national security concerns to engage in various forms of algorithmic authoritarianism.

8.2.2 Not all algorithmic authoritarianism plans succeed: Intended vs real effects of AI authoritarianism

The gap between any intended effects of AI systems for repression and the actual impact is significant for various reasons. Such gaps can result from user resistance, technological shortcomings, or both. This subsection will explore some of the ways intended AI repression practices can fail, leading to less effective surveillance and control in reality.

8.2.2.1 User adaptation and resistance

The phenomenon of **user adaptation and resistance** is a testament to the dynamic interplay between technological imposition and human agency. In authoritarian contexts where regimes seek to harness AI to consolidate their control over information flows, such as the National Information Network (NIN) in Iran, the intended outcomes often rely heavily on citizen compliance and technological efficacy[361]. In Iran, the government's strategy to induce a transition to domestic applications, aiming to facilitate surveillance and

[359] H. A. Ünver, 'Artificial intelligence, authoritarianism and the future of political systems', *Center for Economic and Foreign Policy Research*, 2018.
[360] H. A. Ünver, 'The Role of Technology: New Methods of Information, Manipulation and Disinformation, *Center for Economic and Foreign Policy Research*, 2023.
[361] A. Yalcintas and N.Alizadeh, N., 'Digital protectionism and national planning in the age of the internet: the case of Iran', *Journal of Institutional Economics*, Vol 16, No 4, 2020, pp. 519-536.

censorship, provides a stark illustration of this dynamic[362]. The NIN conceptualised as a parallel to the global internet, is designed to be a controllable network where the state can readily apply AI-driven monitoring tools on domestic applications. It was anticipated that this network would enable comprehensive control over social narratives and restrict the exchange of dissident information, thereby reducing the potential for public dissent and mobilisation.

However, this strategy's practical effectiveness has been questioned by observers such as Akbari and Gabdulhakov (2019), who point out the user resistance to such manoeuvres[363]. Users often exhibit a preference for international applications, such as Telegram, which offer a perception, if not the reality, of better security against government intrusion, a wider range of features and, crucially, access to a global communication network beyond the reach of national surveillance. Resistance can manifest itself in various forms, from a straightforward refusal to adopt state-promoted platforms to more sophisticated methods of bypassing network restrictions[364].

Users may employ virtual private networks (VPNs) to access global internet services, engage in cryptographic communication to obfuscate their online activities or use decentralised platforms to dilute the state's ability to monitor and control[365]. Moreover, this resistance is not static. It evolves as users continually adapt to new levels of state surveillance. They learn from one another, share tactics and develop a communal knowledge base on how to evade state-imposed digital constraints[366]. This collective resilience effectively creates an ongoing challenge to authoritarian aims, demonstrating that while AI and surveillance technologies present new tools for repression, they also ignite a parallel development of innovative counterstrategies among the population.

The case of Iran and the less-than-successful bid to confine the public to a state-controlled digital ecosystem highlights a broader implication: no matter how sophisticated AI-driven surveillance initiatives are, they can often be met with an intrinsic human compulsion to circumvent control and maintain free channels of communication. For policy formulation, recognising this resilience is crucial. Policies aimed at combating digital authoritarianism must not only consider the capabilities of AI systems but also the creative and adaptive ways in which people resist them. This understanding reinforces the need for an approach that is not solely reliant on countering technology with technology, but one that also supports the fundamental human drive for autonomy and freedom of expression.

8.2.2.2 Technological evasion

Technological evasion represents a salient challenge to the efficacy of AI-driven content moderation and surveillance systems. Encryption tools add a further layer of complexity to this dynamic. Strong end-to-end encryption, utilised in various communication platforms, ensures that messages are readable only by the sender and the recipient, rendering intercepted communication by third parties – including AI monitoring systems – indecipherable. Even as machine learning algorithms advance, the mathematical robustness of contemporary encryption can keep unauthorised entities at bay, including AI systems[367].

---

[362] M. Michaelsen, 'Far away, so close: Transnational activism, digital surveillance and authoritarian control in Iran', *Surveillance & Society*, Vol 15, Nos 3 and 4, 2017, pp. 465-470.

[363] A. Akbari and R. Gabdulhakov, 'Platform surveillance and resistance in Iran and Russia: The case of Telegram', *Surveillance & Society*, Vol 17, Nos 1 and 2, 2019, pp. 223-231.

[364] A. Kharazi, 'Authoritarian Surveillance: 'A Corona Test', *Surveillance and society*, Vol 19, No 1, 2021.

[365] L. M. Tanczer, R. McConville, and P. Maynard, 'Censorship and surveillance in the digital age: The technological challenges for academics', *Journal of Global Security Studies*, Vol 1, No 4, 2016, pp. 346-355.

[366] M. Michaelsen, 'Far away, so close: Transnational activism, digital surveillance and authoritarian control in Iran', *Surveillance & Society*, Vol 15, Nos 3 and 4, 2017, pp. 465-470.

[367] M. Michaelsen, 'Authoritarian practices in the digital age| transforming threats to power: The International Politics of Authoritarian Internet Control in Iran', *International Journal of Communication*, Vol 12, 2018, p. 21.

Anonymous browsers, such as **Tor**, offer another avenue for evading AI monitoring. These browsers randomise pathways through a distributed network of relays, thereby dispersing a user's digital footprint across numerous nodes. Consequently, the task of constructing a coherent surveillance picture becomes computationally onerous for AI systems, as the aggregation of disparate data points to a single user becomes a probabilistic challenge with a low likelihood of success. These evasion technologies are not static but are continually refined to stay ahead of surveillance methods. Their development is often characterised by rapid iteration cycles, community-driven enhancements and an open-source ethos that enables widespread collaboration. Thus, they can frequently outpace the adaptive algorithms of AI surveillance systems that are designed to detect or block them.

The perpetual cat-and-mouse game between evasion technologies and AI surveillance underscores a fundamental tension: as AI capabilities expand, so too do the strategies to undermine them. This dynamic necessitates a nuanced approach to content moderation and surveillance – one that is predicated on an understanding of both the capabilities and limitations of AI in the face of adaptive technological evasion.

8.2.2.3 Algorithmic inefficiency

The notion of **algorithmic inefficiency** is particularly pertinent when scrutinising the application of AI in the domain of content moderation and surveillance within authoritarian regimes. The actual effectiveness of such AI systems often falls short of their theoretical potential, a discrepancy highlighted by Yang and Roberts (2023)[368]. In practical terms, AI-driven content moderation systems are tasked with processing and analysing vast amounts of data – a volume that continues to expand exponentially with the proliferation of digital content.

The task of identifying and categorising content based on context, potential infractions, or security threats is compounded by the dynamic nature of human language and the subtleties of cultural and situational context. Algorithms must decipher not only the semantic content of text but also the intent behind it, which can be obscured by irony, satire, or local vernacular. Linguistic nuances present a formidable challenge to AI systems, which often rely on pattern recognition and lack the inherent understanding of language that comes naturally to people. This limitation can lead to both false positives – where benign content is flagged as inappropriate – and false negatives, where problematic content escapes detection.

Moreover, the effectiveness of AI moderation systems can be compromised by their training data. If the data does not fully represent the diversity of languages, dialects and colloquial expressions used across different regions and communities, the AI's ability to moderate content accurately in those contexts is inherently limited[369]. In authoritarian countries, where control of information is a priority, these inefficiencies can be particularly problematic. The state's expectation for AI to serve as a reliable tool for censorship and surveillance collides with the technological reality of AI's current capabilities.

AI systems can become overwhelmed when faced with the contextual complexity of human communication, leading to moderation that is not only inaccurate but often ineffective at fulfilling the intended repressive objectives. This algorithmic inefficiency implies a critical gap between the aspirations of authoritarian regimes to implement pervasive and precise digital control versus the technological limitations of AI. The misalignment of AI's practical utility in the face of complex, human-driven communication ecosystems underscores a significant, albeit often overlooked, aspect of the discourse on AI authoritarianism.

The efficacy of AI systems in content moderation and surveillance is contingent upon their capacity to discern and adapt to the intricate tapestry of **cultural and social norms** that pervade a given society. However, a pervasive challenge emerges from the AI's intrinsic limitations in apprehending the full

[368] E. Yang and M. E. Roberts, 'The Authoritarian Data Problem', *Journal of Democracy*, Vol 34, No 4, 2023, pp. 141-150.
[369] J. Zeng, 'Artificial intelligence and China's authoritarian governance', *International Affairs*, Vol 96, No 6, 2020, pp. 1441-1459.

spectrum of social and cultural subtleties. These limitations are not merely technical but also reflect a gap in AI's ability to process and interpret human context. When AI systems are deployed in the arena of content moderation, they are required to navigate a complex landscape where nuances of speech, historical context and cultural references are paramount[370].

Consequently, the application of AI in such contexts often erodes user confidence in the system's judgments, with people becoming increasingly sceptical of its ability to discern content accurately. As a result, this erosion of trust precipitates a behavioural shift among users, who may seek alternative platforms less prone to such errors, or who may resort to using coded language and other obfuscation techniques to bypass the AI's scrutiny. In essence, the AI's deficiencies in cultural and social understanding create a reactive landscape where users continually evolve their communication strategies in defiance of AI-imposed restraints. This adaptive dynamic not only undermines the intended function of AI systems in authoritative surveillance and moderation but also highlights the inherent challenges in developing AI that can truly comprehend the depth and breadth of human cultural expression.

8.2.2.4 The over-reliance on technological solutions

In the architecture of state-level surveillance and control mechanisms, there has been an emergent trend toward the overestimation of AI's capabilities, which may lead to an over-reliance on technological solutions to the detriment of traditional intelligence and policing methodologies[371]. This inclination to depend heavily on AI for content moderation, surveillance and social management arises from the perception of AI as a panacea that can offer seamless, omnipresent oversight. However, this perspective is often myopic, neglecting the multifaceted nature of control and repression. The deployment of AI systems as primary tools for monitoring and repression is predicated on their ability to analyse large data sets, recognise patterns and make determinations at a scale far beyond human capacity. Nevertheless, the intrinsic limitations of current AI technologies mean that these systems are unable to replicate fully the nuanced understanding and adaptive capabilities of human intelligence[372]. The shortfall is particularly evident in scenarios that require context-sensitive judgements, which are still better handled by trained human operatives. This overreliance on AI can lead to systemic vulnerabilities.

Governments may underinvest in human intelligence assets, both regarding personnel and in cultivating the analytical skills necessary to interpret the complex social, political and cultural landscapes within which control mechanisms must operate[373]. Traditional policing methods, which offer tangible, on-the-ground presence and can provide immediate responses to fluid situations, may be similarly undervalued. In environments where such overreliance exists, certain sections of the populace – particularly those who are technologically literate – can find and exploit gaps in the AI's surveillance net. These individuals may employ a variety of countermeasures, ranging from the use of encrypted communication channels to sophisticated techniques for anonymising online activities, effectively rendering them invisible to AI monitoring.

They may also utilise knowledge of the AI system's specific weaknesses, such as its inability to process ambiguous or deliberately misleading information. These exploits are not merely incidental; they reflect a

---

[370] W. L. Johnson, and A. Valente, 'Tactical language and culture training systems: Using AI to teach foreign languages and cultures', *AI magazine*, Vol 30, No 2, 2009, pp. 72-72.

[371] H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, and R. Krishna, 'Explanations can reduce overreliance on ai systems during decision-making', *Proceedings of the ACM on Human-Computer Interaction*, Vol 7, No CSCW1, 2023, pp. 1-38.

[372] Z. Buçinca, M. B. Malaya, and K. Z. Gajos, K. Z., 'To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making', *Proceedings of the ACM on Human-Computer Interaction at Cornell University*, Vol 5, No CSCW1, 2021, pp. 1-21.

[373] D. F. Engstrom, D. E. Ho, C. M. Sharkey, and M. F. Cuéllar, 'Government by algorithm: Artificial intelligence in federal administrative agencies', *NYU School of Law, Public Law Research Paper*, 2020, pp. 20-54.

deeper systemic issue where the perception of AI as a comprehensive solution overlooks the technology's current developmental stage and its integration into a broader security apparatus. The result is a security paradigm that, while technologically advanced, is characterised by a brittle rigidity and a lack of adaptive resilience that more balanced approaches, incorporating both technological and human elements, can provide. In this way, an overreliance on technological solutions for repression and control engenders a false sense of security, fostering vulnerabilities that, when leveraged, can undercut the efficacy of AI systems and the authoritarian aims they serve.

In essence, while AI offers authoritarian regimes unprecedented capabilities for control and surveillance, the complex interplay of human behaviour, technology and society often renders these systems less effective than intended. Policy-makers must consider these practical limitations when formulating strategies to counter AI-enabled repression, ensuring that measures are holistic and account for the adaptive nature of both technology and society.

8.2.3 Impact of AI technologies on freedoms and rights

The immediate impact of censorship, propaganda and surveillance on freedom of expression as well as other civil and political rights can be profound and multifaceted. Technically, the integration of AI into these activities enhances their scope, precision and surreptitious nature, resulting in a more pervasive infringement on rights. In the realm of censorship, AI algorithms can process vast quantities of data at an unprecedented speed, enabling real-time monitoring and filtering of online content. State actors, using sophisticated machine learning models, can identify and suppress speech they deem undesirable, often with alarming accuracy. NLP technologies allow for the analysis of sentiment, context and even implied meanings within the text, effectively silencing nuanced discourse.

Such algorithmic censorship extends beyond the digital sphere, as AI-driven content moderation systems influence the public sphere by shaping the available information, thus undermining the fundamental right to access information[374]. Propaganda has been revolutionised by AI through the creation of tailored content that targets individuals based on their digital profiles. AI systems can analyse personal data and online behaviour to identify psychological vulnerabilities, allowing state actors to disseminate personalised propaganda. This not only skews the public's perception of reality but also erodes trust in the media and institutions by fostering polarised echo chambers that undermine democratic discourse[375].

Surveillance powered by AI significantly impacts civil liberties, as state actors deploy facial recognition, gait analysis and predictive policing algorithms that infringe on the right to privacy and can lead to arbitrary detention. These systems often operate under the guise of public security, yet they frequently lack transparency and accountability, leaving little room for redress or consent. Furthermore, AI can aggregate data from multiple sources to create detailed profiles of individuals, amounting to a form of digital panopticon that stifles freedom of expression through the chilling effect of perceived omnipresent monitoring[376].

The technical sophistication of AI in these domains presents unique challenges for safeguarding civil and political rights. Traditional methods of protecting privacy and freedom of expression are often ill-equipped to counteract the invasive nature of these technologies. The complexity and opacity of machine learning algorithms, combined with the difficulty in auditing and interpreting their decision-making processes,

[374] E. Aizenberg and J. Van Den Hoven, 'Designing for human rights in AI', *Big Data & Society*, Vol 7, No 2, 2020.
[375] G. Bolsover and P. Howard, 'Computational propaganda and political big data: Moving toward a more critical research agenda', *Big data*, Vol 5, No 4, 2017, pp. 273-276.
[376] F. A. Raso, H. Hilligoss, V. Krishnamurthy, C. Bavitz, and L. Kim, 'Artificial intelligence & human rights: Opportunities & risks', *Berkman Klein Center Research Publication*, 2018.

create an accountability gap. To address these impacts, it is imperative to develop technical standards and regulatory frameworks that ensure transparency, accuracy and fairness in AI applications.

Privacy-preserving technologies that are not directly the topics of this IDA like **differential privacy, federated learning,** and **homomorphic encryption**[377] could be harnessed to mitigate the risks of mass surveillance and data misuse. Furthermore, robust anonymisation techniques and data protection measures need to be integrated into AI systems to safeguard personal information from being exploited for censorship or propaganda. Technical countermeasures, such as adversarial attacks that expose the weaknesses of AI systems, can be a double-edged sword, potentially being used to evade justifiable regulation while also offering a means to resist unjust censorship[378]. These are important concepts and measures which the EU can deploy as part of its export controls and market access for international technology companies.

The proliferation of AI technologies also poses significant risks to physical integrity rights, which encompass protections against violations such as political imprisonment, torture and extrajudicial killings. While AI itself is a neutral technology, its application can have dire consequences when employed within frameworks that disregard human rights.

8.2.3.1 Political Imprisonment

AI technologies enhance the capabilities of authoritarian regimes to monitor and profile political dissidents. With advanced data analytics, governments can sift through massive datasets to identify individuals who might pose a threat to the *status quo*. AI-powered surveillance systems, including facial recognition and social media monitoring, can flag activists, opposition members, or any individuals engaged in protest activities, leading to pre-emptive arrests and political imprisonment. The precision and pervasiveness of these tools mean that the net cast for potential political prisoners is wider and more discriminatory, often based on predictive policing models that operate on biased or speculative data.

8.2.3.2 Torture

Although AI itself does not engage in physical acts of violence, it can facilitate such violations by optimising interrogation schedules or by monitoring the health of detainees to ensure they remain conscious during torture sessions. AI systems could potentially be used to analyse the responses of prisoners to different forms of torture, learning to maximise psychological or physical stress. There is also the psychological dimension, where AI could contribute to the design of interrogation protocols tailored to break down a detainee's resistance. The potential for AI to be misused in such a manner raises significant ethical concerns and strict regulations are required to prevent such applications.

---

[377] **Differential privacy** is a system for publicly sharing information about a dataset by describing patterns of groups within the dataset while withholding information about individuals in the dataset. It provides a way to maximise the accuracy of queries from statistical databases while minimising the chances of identifying its entries. This approach ensures that the removal or addition of a single database item does not significantly affect the outcome of any analysis, thereby protecting the privacy of individuals' data. **Federated learning** on the other hand, is a machine learning approach where a model is trained across multiple decentralised devices or servers holding local data samples, without exchanging them. This process allows for collaborative model training while keeping all the training data local, thus preserving privacy. The central server coordinates the process, aggregating and updating the global model based on locally computed updates. This technique is especially useful for scenarios where data privacy is paramount or where data cannot be centralised due to its size or other constraints. Finally, **homomorphic encryption** is a form of encryption that allows computation on ciphertexts, generating an encrypted result which, when decrypted, matches the result of operations performed on the plaintext. This means it is possible to perform operations on encrypted data without needing to decrypt it first, maintaining data privacy throughout the process. This technique is particularly useful for secure data processing in cloud computing and for maintaining the confidentiality of sensitive data during computation.

[378] H. Fang and Q. Qian,' Privacy preserving machine learning with homomorphic encryption and federated learning', *Future Internet*, Vol 13, No 4, 2021, p. 94.

8.2.3.3 Killings

Perhaps the most alarming potential use of AI in violating physical integrity rights is in the development of lethal autonomous weapons systems. These systems can identify, target and engage without human intervention. The use of such weapons raises profound ethical, legal and security concerns, particularly regarding accountability for AI-driven decisions that result in wrongful death. In conflict zones, AI-driven drones could carry out targeted killings, sometimes based on data that may be inaccurate or misinterpreted by algorithms. The risk of errors leading to civilian casualties is significant, especially without adequate human oversight.

The role of AI in enabling these violations is not just a theoretical concern but a practical one, as the technology could be used to scale and refine oppressive practices. The use of AI in such contexts exacerbates the challenges of holding perpetrators accountable, given the difficulties in attributing decisions made by autonomous systems to individuals or entities. To counter these risks, a robust framework for the ethical development and use of AI is essential. This includes international treaties similar to those that ban other forms of inhumane weaponry, rigorous oversight mechanisms, and the implementation of 'human-in-the-loop' systems to ensure that life-and-death decisions are subject to human judgment and accountability.

It is also important to distinguish between algorithmic authoritarianism cases where there has been evidence, versus hypothetical realistic scenarios. Some of the impacts have been systematically documented, while others are speculative or potential, with limited evidence available to date.

Areas with systematic evidence:

- **Surveillance and privacy:** Systematic evidence exists concerning the impact of AI on privacy rights, as documented by Gohdes (2018; 2020) and others. AI enhances the capacity for mass surveillance, including the monitoring of digital communications and social media. The integration of AI in surveillance technologies allows for the collection, processing and analysis of large datasets on individuals' behaviour patterns, movements and networks, often without consent or adequate data protection, infringing the right to privacy[379].
- **Censorship and freedom of expression:** Studies such as those by Xu (2021) highlight the role of AI in online censorship and content moderation. Automated algorithms can filter, block, or take down content, impacting freedom of expression. While intended to remove harmful content, these AI systems can be overzealous or biased, suppressing legitimate speech[380].
- **Bias and discrimination:** AI systems can perpetuate and amplify biases, leading to discrimination in various sectors, including employment, law enforcement and credit scoring. There is substantial evidence proving that AI can entrench socio-economic disparities by reflecting the prejudices present in their training data or design.

[379] A. R. Gohdes, 'Repression technology: Internet accessibility and state violence', *American Journal of Political Science*, Vol 64 No 3, 2020, pp. 488-503; J. Earl, T. V. Maher, and J. Pan, 'The digital repression of social movements, protest, and activism: A synthetic review', *Science Advances*, Vol 8, No 10, 2022; T. Dragu and Y. Lupu, 'Digital authoritarianism and the future of human rights', *International Organization*, Vol 75 No 4, 2021, pp. 991-1017; M. E. Roberts, 'Resilience to online censorship', *Annual Review of Political Science*, Vol *23*, pp. 401-419; A. R. Gohdes, 'Studying the Internet and Violent Conflict', *Conflict Management and Peace Science*, Vol 5, No 1, 2018, pp. 89-106.

[380] X. Xu, 'To repress or to co-opt? Authoritarian control in the age of digital surveillance' *American Journal of Political Science*, Vol 65 No 2, 2021, pp. 309-325; X. Xu, G. Kostka, and X. Cao, 'Information control and public support for social credit systems in China', *The Journal of Politics*, Vol 84, No 4, 2022, pp. 2230-2245; E. Keremoğlu and N. B. Weidmann, 'How dictators control the internet: A review essay', *Comparative Political Studies*, Vol 53 No 10-11, 2020, pp. 1690-1703.

Areas with potential impact but limited systematic evidence:

- **Autonomous weapons and right to life:** The potential for AI to impact the right to life through autonomous weapons is a growing concern. There is ongoing debate and speculative discussion on the use of lethal autonomous weapons systems in armed conflict, but systematic evidence of their impact on human rights is not yet extensive due to their emerging nature.
- **AI in the criminal justice system:** AI's role in predictive policing and sentencing has the potential to impact human rights related to the administration of justice. While there are some documented cases and studies, the broader systematic impact of AI in the criminal justice system, especially regarding due process and the right to a fair trial, is not fully understood.
- **Social and economic rights:** The potential of AI to affect social and economic rights, such as the right to social security and the right to work, is another area with more speculative than systematic evidence. AI's role in automating jobs has implications for employment and the broader economy, but comprehensive, systematic evidence of these impacts is still developing.
- **Access to information:** AI systems that curate and recommend content can influence individuals' access to information. Although there are individual cases and concerns, systematic evidence regarding the broader impact of AI on this aspect of human rights is still limited.
- **Psychological well-being:** There is growing concern about AI's potential impact on psychological well-being, particularly through social media algorithms that may contribute to addiction or mental health issues. However, systematic evidence directly linking AI to these outcomes is still emerging.

As AI technologies continue to evolve, the evidence base for their impact on human rights will grow. It is essential for researchers, policy-makers and advocates to monitor these developments, conduct thorough investigations and advocate for the responsible use of AI to safeguard human rights. The gaps in systematic evidence highlight the need for continued empirical research and the establishment of mechanisms to assess and mitigate potential human rights violations proactively.

PE 754.450
EP/EXPO/DROI/FWC/2019-01/Lot6/1/C/30

Print ISBN 978-92-848-1856-3 | doi: 10.2861/907329 | QA-02-24-534-EN-C
PDF ISBN 978-92-848-1855-6 | doi: 10.2861/52162 | QA-02-24-534-EN-N