IEAI Technical University of Munich
School of Social Sciences and Technology
Institute for Ethics in Artificial Intelligence

Research Brief – Q3 2026

Autonomy Lost? Human Rights as "Negative Obligations" in the Age of AI

By Alexander Kriebitz and Alberto Santos Ferro

The growing integration of artificial intelligence (AI) in society has reinforced a shift from abstract ethical principles to human rights as the underlying normative basis for AI governance. While there is an emerging consensus that AI development, deployment and use should be guided by international human rights law, its exact implications for the use of AI by state actors remain contested.

This Research Brief focuses on the function of human rights as negative obligations of states that constrain the exercise of power by states over individuals. The authors argue that AI poses a structural challenge to this protective function of human rights, particularly when it comes to rights associated with human self-expression and freedom of thought. To address these developments, the Brief proposes the doctrinal expansion of existing human rights to protect individual autonomy against the potential misuse of AI by state actors.

Introduction

Artificial intelligence (AI) is increasingly being used in areas that affect personal autonomy, self-expression and freedom of thought: AI systems can access and process intimate personal data and increasingly infer individuals' mental states (Tortora, 2020; Oravec, 2022; Peters & Matz, 2024). At the same time, researchers are increasingly studying how LLMs affect formation of individual and collective opinions (Li et al., 2023). Such developments raise questions about the potential misuse of AI, for instance, when it comes to the use of neuroprediction technologies to predict criminal behavior (see Tortora et al., 2020).

While high-level principles increasingly position human rights as a normative basis for AI governance (see UN General Assembly, 2024), the interpretation of constitutional and international human rights law is often marked by normative tensions, for instance between restrictions on freedom of movement during pandemics and between freedom of expression and state obligations to prevent discrimination (Zucca, 2007; (Kaye, 2017; Fontes et al., 2022).

For this reason, this Research Brief examines one specific dimension of human rights; their interpretation as negative obligations constraining the use of state power in areas concerning individual autonomy, freedom of thought and self-determination (Berlin, 1969). Its central claim is that AI poses a structural challenge to these obligations as AI systems increasingly influence the formation and expression of thoughts, utilize intimate personal data and shape access to information through algorithmic curation. In response, the brief argues for a doctrinal expansion of existing human rights protections and proposes strengthening the role of human rights as red lines for the use of AI by the state.

Human Rights as Negative Obligations

While covering a broad range of different rights, human rights encapsulate the view that state actors do not possess absolute authority over their citizens (Murray, 2016). In order to address situations of conflicting interpretations of human rights, legal theories typically distinguish negative obligations, which require states to refrain from unjustified interference with protected rights, and positive obligations, which require states to take affirmative measures to secure their effective enjoyment (Mégret, 2010).

AI poses a structural challenge…as AI systems increasingly influence the formation and expression of thoughts, utilize intimate personal data and shape access to information through algorithmic curation.

Human Rights as Constraints for State Actors

For much of human history, political orders were characterized by dense social control, strong expectations of conformity and limited access to positions of power and prestige (Black, 2011). Against this backdrop, the emergence of rights as safeguards against governmental interference represents a profound innovation in political thought. It transformed not only the understanding of human beings as persons with inherent moral status but also limited the exercise of state power over its citizens (Hughes, 2025). Rather than empowering authorities to enforce paternalistic conceptions of popular virtues, modern constitutionalism assumes that reasonable people may disagree about what constitutes a good life (Mill, 1859; Black, 2011).

Ultimately, these philosophical ideas contributed to the articulation of rights that protect individuals against state overreach (see Berlin, 1969). In German constitutional law, for instance, this is reflected in the understanding of basic rights as "defense rights of the individual" against state interference (Bundesverfassungsgericht, 15 January 1958). Constitutional rights and international human rights thus share a common function of constraining state power, although they do so from different legal perspectives: constitutional rights primarily constrain state action within the domestic legal order, while international human rights impose obligations on states vis-à-vis the international community (D'Amato, 2017).

Source Title Page Image: vecteezy.com, Oleg Gapeenko

Freedom of Conscience and Freedom of Thought

The rights to freedom of conscience and freedom of thought further delineate areas that are inaccessible to societal and political pressure. Their roots lie in medieval canon law, which distinguished between the inner sphere of conscience and personal belief (forum internum) and the public expression of personal views (forum externum) (Roberts, 2020).

Against the backdrop of World War II, this distinction gained renewed political significance as a means of protecting individuals from ideological and religious pressures to conform. The totalitarian regimes of the 20th century demonstrated how easily political ideologies could penetrate individuals' access to information, beliefs, identities and consciences. Literary works such as Orwell's *Nineteen Eighty-Four* captured this danger by portraying a society in which surveillance ultimately destroys the possibility of independent thought (Orwell, 1949). Reflecting such concerns, the concept of the *forum internum* has become an important reference point for European human rights law, and relatedly constitutional law around the globe, particularly in the interpretation of protections for freedom of conscience and freedom of thought (Petkoff, 2012; Benzinger, 2026).

Artificial Intelligence as an Extension of Human Agency

Historically, technological innovations have been associated with new forms of political control (Elden, 2003; Bartlett, 2022). This raises the question of the implications for AI from the perspective of power use in society.

In recent years, a broad scholarly consensus has emerged regarding the normative challenges posed by AI to human oversight, accountability, privacy, fairness and explainability (Floridi et al., 2018). Apart from these wider issues, some normative concerns surrounding AI run deeper. Conventional definitions treat AI as the capacity of machines to perform tasks that require or even outperform biological intelligence (Naeem & Hauser, 2024). The normative issue lies, therefore, not only in the prevention of unintended consequences, for instance, algorithmic discrimination, but also resides in the use of this extended intelligence to realize their preferences more effectively. Concrete examples would be the extension of the "will" of a company to hire employees and to relieve the stress of an HR department combating the flood of incoming (and sometimes AI-generated) applications (see Naeem & Hauser, 2024). This technologically facilitated cognitive extension raises three interrelated normative challenges.

First, AI systems can reflect normative choices made during their design and development, before users ever interact with them (van de Poel, 2020). Autonomous vehicles, for example, require developers to incorporate morally relevant decisions into system architectures before particular situations arise (Geisslinger, Poszler, and Lienkamp, 2023). Such normativity may be deliberate, but can also emerge from technical design choices or optimization objectives, such as engagement-based metrics in recommender systems (van de Poel, 2020). Regardless of intent, these choices can shape human behavior through direct mechanisms, such as predefined options, or more indirect ones, such as nudging or deception.

Secondly, AI increasingly mediates human decision-making across different contexts. AI systems increasingly determine who receives a loan, a job or social benefits or medical treatment. While human operators may formally retain responsibility, they often rely heavily on algorithmic recommendations when making practical decisions (Krügel et al., 2022).

Thirdly, AI is increasingly used for generating knowledge and distributing information. From apps calculating the caloric content of food to radiological analysis, AI systems reproduce and distribute information that informs human decisions. Often perceived as a knowledge-distribution system, AI is increasingly shaping individuals' access to information (Lavazza and Farina, 2026).

The normative issue lies, therefore, not only in the prevention of unintended consequences, for instance, algorithmic discrimination, but also resides in the use of this extended intelligence to realize their preferences more effectively.

The Implications of AI for Negative Human Rights Obligations

AI redefines societal decision-making processes, changes the distribution of knowledge in society and enables the projection of norms in different societal contexts. As an extension of human agency, AI thereby enables state power to reach precisely into domains that negative obligations are intended to shield from state intervention, including private life, thought and conscience. The central concern is therefore not AI as such, but the possibility that states may use or regulate AI in ways that extend their authority into these protected domains.

Two features of state power make its use of AI especially consequential. States are the original duty-bearers under international human rights law, and their power is constrained by constitutions (D'Amato, 2027). Additionally, states possess the monopoly on the legitimate use of force. These very features imply that transgressions of normative standards by states can undermine the very political order that is meant to preserve human rights.

In order to gain a better understanding of the concrete risks associated with the use of AI by states, we examine three interrelated contexts. First, AI can enable states to gain access to increasingly intimate personal data and to derive information about individuals that they have not voluntarily disclosed to public authorities. Second, states may use or regulate AI systems to embed political preferences in the informational environments through which individuals form opinions and beliefs. Third, AI can enable states to make ex-ante assessments of individuals' behavior, potentially turning predictions about future conduct into grounds for state intervention. Across these contexts, the common concern is the expansion of state power into domains in which negative human rights obligations are intended to preserve individual autonomy.

State Access to Intimate Personal Data through Increased AI Deployment

Although personal data has long been used for administrative and commercial purposes, contemporary AI technologies such as large language models (LLMs), conversational chatbots, facial recognition systems and brain-computer interfaces enable unprecedented forms of surveillance by aggregating diverse datasets (see Mozur, 2019; Feldstein, 2019). Data originally collected for therapeutic, administrative or personal purposes may subsequently be used to infer individuals' beliefs, intentions, emotions or mental states, thereby extending the potential reach of state into domains traditionally classified as "intimate personal data" (see Bundesverfassungsgericht, 1983).

Contemporary AI systems already transform the nature of informational privacy. A prominent example is the use of chatbots for personal conversations, journaling, emotional support and other forms of self-disclosure, which contain highly sensitive information about an individual's relationships, health, beliefs or emotional state (Peters, & Matz, 2024; Luo et al., 2025). Such data can, when processed at scale, be used to construct detailed behavioral, psychological and relational profiles (Fan et al., 2023). Recent advances in neurotechnology further intensify these challenges, as they can interpret neural activity and link these to expressions of will and concrete mental states (Ienca, 2023).

AI redefines societal decision-making processes, changes the distribution of knowledge in society and enables the projection of norms in different societal contexts.

Many applications of such AI systems are entirely legitimate. AI-powered brain-machine interfaces, for example, can support patients with locked-in syndrome, while LLMs can be developed to support individuals seeking mental health support (Kübler et al., 2001). The challenge, therefore, lies in preventing the misuse or repurposing of data for political purposes outside of the original context of use. From the perspective of states' negative obligations, the central concern is that state agencies, such as law enforcement or intelligence authorities, gain access to highly intimate data collected by actors in the AI lifecycle for different purposes, potentially without the individual's knowledge or consent. Such access may allow the state to infer information about individuals' beliefs, emotions or mental states, thereby extending state power into domains closely connected to the

forum internum. The use of health records in migration enforcement exemplifies this risk: data collected for legitimate purposes may later be accessed or repurposed by public authorities for purposes that adversely affect the individuals concerned (Center for Health Law and Policy Innovation, 2026, April 22).

Historical precedents illustrate the consequences of such developments. The surveillance apparatus of the German Democratic Republic operated not primarily through extraordinary forms of intrusion, but through the systematic aggregation of individual observations into comprehensive behavioral dossiers (Lewis, 2025). Today's AI systems are vastly more sophisticated than East German bureaucracies: AI-driven biometric surveillance systems, such as facial recognition technologies deployed in public spaces (Feldstein, 2019; Fontes et al., 2022) or lie detectors that infer deception, emotions or intentions from facial micro-expressions, voice characteristics or physiological signals (Oravec, 2022).

More fundamentally, the use of facial recognition tools in the judiciary can challenge defence rights by undermining the right to remain silent. In the United States, the protection from self-incrimination afforded by the Fifth Amendment embodies the principle that individuals cannot be compelled to contribute to their own conviction through coerced disclosure (Amar & Lettow, 1995). The rationale of the mentioned AI systems raises, therefore, broader questions about their permissibility and human rights impacts, but more general the access of state agencies to intimate data collected by companies or public administration, for instance patient health records (Deskus, 2018; Center for Health Law and Policy Innovation, 2026, April 22).

Embedding Political Preferences in AI

LLMs, search engines and recommender systems not only shape what users know, but they also increasingly assume roles traditionally distributed across pluralistic media systems (Li et al., 2023). At the same time, they do not merely present data in a neutral manner, but actively structure what counts as relevant and meaningful information: in the best case, to reduce the complexity of information presented, in the worst case, to actively shape views of individuals (Cheng et al., 2025). While these systems are usually guided by a commercial rationale, legislation and ethical frameworks have increasingly sought to align them with overarching societal preferences. The question concerns therefore the use of state legislation to integrate political and broader societal preferences in AI.

[LLMs, search engines and recommender systems] do not merely present data in a neutral manner, but actively structure what counts as relevant and meaningful information: in the best case, to reduce the complexity of information presented, in the worst case, to actively shape views of individuals (Cheng et al., 2025).

Recent years have seen a growing emphasis on aligning AI systems with specific normative preferences, some of which serve legitimate protective purposes (Huang, Papyshev and Wong, 2025). It is generally uncontroversial, for example, when legislative frameworks such as the Digital Services Act aim to oblige providers of AI systems to address harmful content, such as non-consensual intimate imagery (Lemoine & Vermeulen, 2024).

Greater challenges arise, however, when AI systems are aligned with normative preferences concerning contentious moral, religious or political questions. Issues such as abortion illustrate this difficulty. LLMs exhibit measurable differences in how they explain abortion law and policy across jurisdictions, reflecting both legal and normative choices (Encarnación et al., 2026). Likewise, research on tonality of language in LLMs illustrates the potential of the use of LLMs to reinforce or attenuate feelings of guilt, shame and social stigma through the language used in their responses (Sharma et al., 2025, Okoso et al., 2025). Developments in China and Russia illustrate efforts to align LLMs and recommender systems with political preferences, demonstrating how such systems can be used to promote or reinforce political preferences such as conversative values (Pan & Xu, 2026; The Moscow Times, 2026, January 16).

From the perspective of negative obligations, such forms of alignment raise a distinct concern: the potential overreach of state power into domains closely connected to the formation of thoughts,

beliefs and emotions. Such algorithmic designs may therefore affect not only the right "to seek, receive and impart information" (Article 19, Universal Declaration of Human Rights), but also the *forum externum* dimension of freedom of conscience—understood here as protection from propaganda, ideological pressure and externally imposed belief formation (Article 18, Universal Declaration of Human Rights). Legal frameworks regulating AI in contexts that are capable to influence the formation of opinion should therefore be subjected to more rigorous proportionality assessments that consider potential negative effects on freedom of expression, freedom of conscience and individual autonomy (Kaye, 2017).

Yet history shows that both individuals and institutions are fallible: political convictions, ideological commitments, fears, miscalculations, corruption and group-based stereotypes can shape decision-making within political and legal institutions.

The Use of AI Systems for the Prognosis of Individual Behavior

The further challenge advanced AI models pose to defense rights lies in their use in making judgments about human behavior and in estimating how likely they are to act in certain ways (Loefflad, Chen, & Grossklags, 2024). While the prognosis of human behavior might fulfil important roles, for instance when it comes to forecasting traffic, energy consumption or even the modeling of pandemic situations, the use of ex-ante assessments on an individual basis, for instance AI-powered neuroprediction, faces significant legal hurdles if connected to law enforcement and potentially penal consequences (Tortora et al. 2020).

A relevant example is the guideline against "de-extremification" in Xinjiang, where otherwise innocuous religious and cultural practices such as abstaining from alcohol or smoking, fasting and wearing beards or veils were identified as indicators of "extremification" (Finley, 2018). This illustrates how behavioral practices that are per se harmless can become markers of political suspicion when interpreted through a state-defined framework of extremism. In the context of Xinjiang's re-education policies, it highlights the risk of combining AI with expansive definitions of deviance to enable intrusive forms of behavioral control. (Kam & Clarke, 2021).

Another relevant issue are AI solutions that aim to predict criminal behavior on an individual basis (Blount, 2024). For instance, the Dutch police use AI systems to perform risk evaluations on an individual level (Schuilenburg & Soudijn, 2023). The deployment of such systems raises questions regarding algorithmic bias, but also with respect to overarching anthropological assumptions embedded beneath (Byler, 2019). In addition, the case illustrates that the use of AI in ways conflicting with human rights is not confined to authoritarian regimes.

Negative Obligations under the Condition of Imperfect Institutions

AI ethics can, in certain instances, rely on an idealized view of institutions, legal systems and enforcement mechanisms (see North, Wallis & Weingast, 2009). Yet history shows that both individuals and institutions are fallible: political convictions, ideological commitments, fears, miscalculations, corruption and group-based stereotypes can shape decision-making within political and legal institutions. Cases such as the use of enhanced interrogation techniques at Guantanamo Bay, policing during the Troubles in Northern Ireland, and digital surveillance of migrants in the United States illustrate how even stable democracies may restrict fundamental rights in times of political crisis (Duffy, 2019; Feathers, 2021).

AI does not eliminate these institutional vulnerabilities. Rather, it can amplify them by enabling systemic infringements of individual rights at greater scale. Decisions that might previously have depended on individual officials can increasingly be supported by AI systems, allowing discriminatory criteria to be reproduced across large populations. In contexts of war, unrest or other political crises, this combination of institutional fallibility and technological scale reinforces the need for safeguards that constrain the use of AI on a constitutional basis.

Implications for Rights-based AI Governance

In the following, we explore potential pathways that could address the risks posed by AI in relation to negative human rights obligations that seek to protect individual moral autonomy. The primary objective of the proposed measures is to prevent the use of technological systems such as brain machine interfaces, LLMs or recommender systems and the data generated or collected through their use in ways that replicate historical patterns of political control and interference.

AI-enabled neurotechnology and increasingly sophisticated behavioral inference systems create a qualitatively new challenge to freedom of thought and freedom of conscience.

Protecting the Forum Internum in the Age of Behavioral AI

The *forum internum*, the inner domain of thought, belief and moral deliberation, has traditionally been understood as the most protected layer of individual self-expression. Owing to the increasing proximity of AI-enabled technologies to cognitive processes and the formation of thought, existing protections of the right to freedom of conscience and freedom of thought may require doctrinal expansion. Specifically, we argue for refinement in the following three areas:

a. Explicit prohibition of the criminalization of thought and mental states

AI-enabled neurotechnology and increasingly sophisticated behavioral inference systems create a qualitatively new challenge to freedom of thought and freedom of conscience. Given the historic use of states for surveillance, the perceived ability to infer mental states, intentions or beliefs, creates a risk that such technologies could be used to criminalize particular thoughts or deviant ideological assumptions in ways that were previously impossible from a technical standpoint. Such practices would constitute an inherent interference with the *forum internum.* The negative obligation to respect freedom of thought must therefore entail that states refrain from criminalizing, coercing or otherwise regulating thoughts and mental states themselves.

b. Protection of data revealing mental states

The protection of freedom of thought necessarily extends beyond thoughts themselves to information capable of revealing them. AI increasingly derives intimate psychological and cognitive inferences from behavioral, biometric and potentially neural data. Existing privacy protections should therefore be expanded to recognize data that directly reveals or reliably infers thoughts, intentions, beliefs or mental states as a distinct category of protected information (Council of Europe, 2021; Ienca, 2023).

c. Prohibition of ex-ante behavioral assessments of individuals

Human dignity rests on the premise that individuals are capable of change and that individuals are not permanently defined by past behavior. AI systems that predict an individual's future criminality, dangerousness or other legally relevant conduct challenge this premise by replacing individualized legal judgment with probabilistic classification. Predictive assessments of individual behavior should be prohibited, including judicial authorization, transparent evidentiary standards, meaningful human review, the right to contest algorithmic findings and safeguards against protected characteristics or opaque statistical proxies.

d. Limits on the extension of political preferences through legislation

The use of AI to realize societal preferences also raises questions about the purpose of law. As AI enables the large-scale prediction, classification and optimization of human behavior, it creates political pressure to extend legal regulation beyond harmful conduct towards the promotion of socially desirable behavior. The liberal tradition, which forms the intellectual basis of negative obligations, has traditionally rejected such an expansive role for the state (Mill, 1859; Berlin, 1969). Likewise, the right to freedom of conscience implies that certain decisions must remain within the sphere of individual autonomy, including personal choices relating to abortion, the “inherently private choices”, or acts of civil disobedience (see Supreme Court of Canada, 1997; Brownlee, 2012).

e. Limits on the use of AI in law enforcement arising from human fallibility

A further implication concerns the preservation of human fallibility as an essential feature of autonomy. Apart from the prohibition of ex-ante profiling, recognizing human fallibility as an intrinsic part of the human condition may therefore provide a counterbalance to the increased efficiency of AI-enabled law enforcement, building on earlier legal innovations such as the right to be forgotten as discussed by Mayer-Schönberger (2010). The underlying concern is thus not merely to prevent the optimization of law-enforcement efficiency, but to preserve a human-rights-based legal order in which individuals retain the possibility of change or rehabilitation.

Conclusion

This Brief argues that AI poses a structural challenge to the function of human rights as constraints for the exercise of power by the state. By expanding the ability to collect intimate personal data, infer mental states, predict behavior and shape decision-making, AI reaches into domains that constitutional and human rights have traditionally sought to place beyond political authority.

Human rights understood as negative obligations provide a principled response to this tendency by reaffirming that the purpose of law is not to perfect human beings, but to preserve the conditions under which individuals can think, choose, err and develop autonomously.

On this basis, the authors argue for a doctrinal expansion of negative obligations in five interconnected areas to protect individuals from state overreach:

- the explicit prohibition on the criminalization of thought, emotions and other aspects of the inner life of persons through constitutional safeguards;
- the heightened protection of intimate personal data from access by state actors;
- the prohibition of ex-ante behavioral assessments of individuals by state actors;
- the limits on the extension of political preferences through legislation; and
- the limits on the use of AI in law enforcement arising from human fallibility.

These proposals are not intended to create new human rights, but to preserve the protective function of existing ones under changing technological conditions.

References

Amar, A. R., & Lettow, R. B. (1995). Fifth amendment first principles: The self-incrimination clause. *Michigan Law Review*, *93*(5), 857–928.

Bartlett, R. (2022). *The Bentham Brothers and Russia: The imperial Russian constitution and the St Petersburg panopticon*. UCL Press.

Benziger, F. (2026). The Concept of the Inner Self in European Human Rights Law. *European Journal of Comparative Law and Governance*, *1*(aop), 1-36.

Berlin, I. (1969). Two concepts of liberty. In *Four essays on liberty* (pp. 118–172). Oxford University Press.

Black, D. (2011). *Moral time*. Oxford University Press.

Blount, K. (2024). Using artificial intelligence to prevent crime: Implications for due process and criminal justice. *AI & SOCIETY*, *39*(1), 359–368.

Brownlee, K. (2012). *Conscience and conviction: The case for civil disobedience*. OUP Oxford.

Bundesverfassungsgericht [Federal Constitutional Court]. (1958, January 15). *Urteil des Ersten Senats – 1 BvR 400/51 (Lüth)*.

Bundesverfassungsgericht. (1983, December 15). *BVerfGE 65, 1 (Volkszählung)*. Das Fallrecht. https://www.servat.unibe.ch/dfr/bv065001.htm

Byler, D. (2019). Preventative policing as community detention in Northwest China. *Made in China Journal*, *4*(3).

Center for Health Law and Policy Innovation. (2026, April 22). *Medicaid data sharing with ICE: What we know and why we're worried*. https://chlpi.org/wp-content/uploads/2026/04/Medicaid-ICE-data-sharing-FINAL.pdf

Cheng, M., Ofek, E., & Yoganarasimhan, H. (2025). Balancing engagement and polarization: Multi-objective alignment of news content using LLMs. *arXiv preprint arXiv:2504.13444*.

China Law Translate. (2017, March 30). *Xinjiang Uyghur Autonomous Region Regulation on De-extremification*. https://www.chinalawtranslate.com/en/xinjiang-uyghur-autonomous-region-regulation-on-de-extremification/

Coeckelbergh, M. (2026). AI and epistemic agency: How AI influences belief revision and its normative implications. *Social Epistemology*, *40*(1), 59–71.

Council of Europe. (2021). *Feasibility study on a legal framework for AI*. Council of Europe.

D'Amato, A. (2017). The concept of human rights in international law. In *International law of human rights* (pp. 21–70). Routledge.

Deskus, C. (2018). Fifth amendment limitations on criminal algorithmic decision-making. *New York University Journal of Legislation and Public Policy*, *21*, 237–268.

Duffy, A. (2019). *Torture and human rights in Northern Ireland: Interrogation in depth*. Routledge.

Elden, S. (2003). Plague, panopticon, police. *Surveillance & Society*, *1*(3), 240–253.

Encarnación, R., Jones, C., & Metaxa, D. (2026, June). Auditing LLM Responses in a Complex Policy Landscape: Abortion Law in the United States. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency* (pp. 7237-7262).

Fan, J., Sun, T., Liu, J., Zhao, T., Zhang, B., Chen, Z., ... & Hack, E. (2023). How well can an AI chatbot infer personality? Examining psychometric properties of machine-inferred personality scores. *Journal of Applied Psychology*, *108*(8), 1277.

Feathers, T. (2021, March 4). 'They track every move': How US parole apps created digital prisoners. *The Guardian*. https://www.theguardian.com/global-development/2021/mar/04/they-track-every-move-how-us-parole-apps-created-digital-prisoners

Feldstein, S. (2019). *The global expansion of AI surveillance*. Carnegie Endowment for International Peace. https://carnegieendowment.org/research/2019/09/the-global-expansion-of-ai-surveillance

Floridi, L., Cowley, J., Beltrametti, M., Chatila, R., Dignum, V., Emanuel, E., ... & van Wynsberghe, A. (2018). AI4People—an ethical framework for a good AI society. *Minds and Machines*, *28*(4), 689–707.

Fontes, C., Hohma, E., Corrigan, C. C., & Lütge, C. (2022). AI-powered public surveillance systems: Why we (might) need them and how we want them. *Technology in Society*, *71*, 102137.

Geisslinger, M., Poszler, F., & Lienkamp, M. (2023). An ethical trajectory planning algorithm for autonomous vehicles. *Nature Machine Intelligence*, *5*(2), 137–144.

Hayek, F. A. (1960). *The constitution of liberty*. University of Chicago Press.

Huang, L. T. L., Papyshev, G., & Wong, J. K. (2025). Democratizing value alignment: From authoritarian to democratic AI ethics. *AI and Ethics*, *5*(1), 11-18.Casademunt, H., Tran, K., Jakkli, A., Marks, S., & Nanda, N. (2026). Censored LLMs as a natural testbed for secret knowledge elicitation. *arXiv preprint arXiv:2603.05494*.

Hughes, G. (2025). *Inherent human dignity: A philosophical meditation*. University of Notre Dame Press.

Ienca, M. (2023). On artificial intelligence and manipulation. *Topoi*, *42*(3), 833–842.

Isaak, J., & Hanna, M. J. (2018). User data privacy: Facebook, Cambridge Analytica, and privacy protection. *Computer*, *51*(8), 56–59.

Kam, S., & Clarke, M. (2021). Securitization, surveillance and 'de-extremization' in Xinjiang. *International Affairs*, *97*(3), 625–642.

Kim, T., Shin, D., Kim, Y. H., & Hong, H. (2024, May). DiaryMate: Understanding user perceptions and experience in human-AI collaboration for personal journaling. In *Proceedings of the 2024 CHI conference on human factors in computing systems* (pp. 1-15).

Krügel, S., Ostermaier, A., & Uhl, M. (2022). Zombies in the loop? Humans trust untrustworthy AI-advisors for ethical decisions. *Philosophy & Technology*, *35*(1), 17–39.

Kübler, A., Kotchoubey, B., Kaiser, J., Wolpaw, J. R., & Birbaumer, N. (2001). Brain–computer communication: Unlocking the locked in. *Psychological bulletin*, *127*(3), 358.

Lemoine, L., & Vermeulen, M. (2024). *Assessing the extent to which generative artificial intelligence (AI) falls within the scope of the EU's Digital Services Act: An initial analysis*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4702422

Lewis, A. (2025). Collaborating with the East German Secret Police: Reading the Stasi Archives with Counter-Archives. *East Central Europe*, *52*(2–3), 191–217.

Li, C., Su, X., Han, H., Xue, C., Zheng, C., & Fan, C. (2023). Quantifying the impact of large language models on collective opinion dynamics. *arXiv preprint arXiv:2308.03313*.

Loefflad, C., Chen, M., & Grossklags, J. (2024). Social scoring systems for behavioral regulation: An experiment on the role of transparency in determining perceptions and behaviors. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (Vol. 7, pp. 891–904).

Luo, X., Wang, Z., Tilley, J. L., Balarajan, S., Bassey, U. A., & Cheang, C. I. (2025). Seeking Emotional and Mental Health Support From Generative AI: Mixed-Methods Study of ChatGPT User Experiences. *JMIR Mental Health*, *12*(1), e77951.

Mégret, F. (2010). The nature of international human rights obligations. In M. Selcuk & S. O. A. Mervan (Eds.), *International human rights law* (pp. 160–180). IIUM Press. https://doi.org/10.31436/iiumlj.v31i2.856

Mill, J. S. (1859). *On liberty*. John W. Parker and Son.

Mozur, P. (2019). *Face scans in China*. *The New York Times*. https://www.nytimes.com

Murray, R. (2016). Absolute and relative rights in international human rights law: A critical analysis. *Human Rights Law Review*, *16*(1), 1–29.

Naeem, H., & Hauser, J. (2024). Should we discourage AI extension? Epistemic responsibility and AI. *Philosophy & Technology*, *37*(3), 91.

North, D. C., Wallis, J. J., & Weingast, B. R. (2009). Violence and the rise of open-access orders. *Journal of Democracy*, *20*(1), 55–68.

Kaye, D. (2018). *Report on artificial intelligence technologies and implications for freedom of expression and the information environment*. OHCHR. https://www.ohchr.org/en/calls-for-input/report-artificial-intelligence-technologies-and-implications-freedom-expression-and

Okoso, A., Yang, M., & Baba, Y. (2025). *Do expressions change decisions? Exploring the impact of AI's explanation tone on decision-making*. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Article 824, pp. 1–22). Association for Computing Machinery. https://doi.org/10.1145/3706598.3713744

Oravec, J. A. (2022). The emergence of "truth machines"?: Artificial intelligence approaches to lie detection. *Ethics and Information Technology*, *24*(1), 6–15.

Orwell, G. (1949). *Nineteen eighty-four*. Secker & Warburg.

Pan, J., & Xu, X. (2026). Political censorship in large language models originating from China. *PNAS Nexus*, *5*(2), pgag013.

Peters, H., & Matz, S. C. (2024). Large language models can infer psychological dispositions of social media users. *PNAS Nexus, 3*(6), pgae231. https://doi.org/10.1093/pnasnexus/pgae231

Roberts, C. K. (2020). *Reconceptualising the place of the forum internum and forum externum in Article 9 of the European Convention on Human Rights* (Doctoral dissertation, University of Bristol). University of Bristol Research Portal.

Schuilenburg, M., & Soudijn, M. (2023). Big data policing: The use of big data and algorithms by the Netherlands Police. *Policing: A Journal of Policy and Practice*, *17*, paad061.

Sharma, A., Mampally, M., Ravuru, C., Brennan, K., & Gaikwad, N. (2025). Can LLMs Understand What We Cannot Say? Measuring Multilevel Alignment Through Abortion Stigma Across Cognitive, Interpersonal, and Structural Levels. *arXiv preprint arXiv:2512.13142*.

Supreme Court of Canada. (1997). *Godbout v. Longueuil (City)*, 1997 3 SCR 844. https://scc-csc.lexum.com/scc-csc/scc-csc/en/item/1560/index.do

Lavazza, A., & Farina, M. (2026). Generative AI as a knowledge distribution system. *Social Epistemology*, 1-17.

The Moscow Times. (2026, January 16). *В России призвали срочно создать православный искусственный интеллект*. https://ru.themoscowtimes.com/2026/01/16/v-rossii-prizvali-srochno-sozdat-pravoslavnii-iskusstvennii-intellekt-a184724

Tortora, L., Meynen, G., Bijlsma, J., Tronci, E., & Ferracuti, S. (2020). Neuroprediction and AI in forensic psychiatry and criminal justice: A neurolaw perspective. *Frontiers in Psychology*, *11*, 220.

United Nations General Assembly. (2024). *Seizing the opportunities of safe, secure, and trustworthy artificial intelligence systems for sustainable development* (Resolution A/78/L.49).

van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds and Machines*, *30*(3), 385–409. https://doi.org/10.1007/s11023-020-09537-4

Zucca, L. (2007). *Constitutional dilemmas: Conflicts of fundamental legal rights in Europe and the USA*. Oxford University Press.