Toward Resisting AI-Enabled Authoritarianism

Fazl Barez
University of Oxford
Whitebox

Isaac Friend
University of Oxford

Keir Reid
Independent Researcher

Igor Krawczuk
Independent Researcher

Vincent Wang
University of Oxford

Jakob Mökander
Tony Blair Institute

Philip Torr
University of Oxford

Julia Morse
University of Oxford

Robert Trager
University of Oxford

Abstract

Artificial-intelligence systems built with statistical machine learning have become the operating system of contemporary surveillance and information control, spanning both physical and online spaces. City-scale face-recognition grids, real-time social-media takedown engines and predictive "pre-crime" dashboards share four politically relevant technical features: massive data ingestion, black-box inference, automated decision-making, and no human in the loop. These features now amplify authoritarian power and erode liberal-democratic norms across many political regimes. Yet mainstream machine learning research still devotes only limited attention to technical safeguards such as differential privacy, federated-learning security and large-model interpretability, or adversarial methods that can help the public resist AI-enhanced domination. We identify four resulting gaps: *evidence* (little empirical measurement of safeguard deployment), *capability* (open problems such as billion-parameter privacy–utility trade-offs, causal explanations for multimodal models and Byzantine-resilient federated learning), *deployment* (public-sector AI systems almost never ship with safeguards enabled by default) and *asymmetry* (authoritarian actors already enjoy a "power surplus," so even incremental defensive advances matter). We propose re-directing the field toward a triad of safeguards—privacy preservation, formal interpretability and adversarial user tooling—and outline concrete research directions that fit within standard ML practice. Shifting community priorities toward *Explainable-by-Design, Privacy-by-Default* is a pre-condition for any durable defense of liberal democracy.

1 Introduction

The development of statistical learning-driven AI systems enables new degrees and new forms of social control. Contemporary AI systems differ from previous technologies in four ways relevant to political power: their ability to process unprecedented volumes of data in real time, their expanded capacity for automated decision-making without human intervention, their predictive modeling, and their black-box nature. Meanwhile, the global social context in which technology must be considered is characterized in part by an increase in authoritarian politics, enough that it makes sense to speak of an international trend. [1] In this paper, we do not assert (nor do we deny) a central causal role for ML technology in this major international political change, but establish by example that **development**

[1] For example, the 2024 Varieties of Democracy Report, which uses a Liberal Democracy Index to rate countries' political regimes, has identified a "wave of autocratization" encompassing 42 countries, many of

Preprint. Under review.

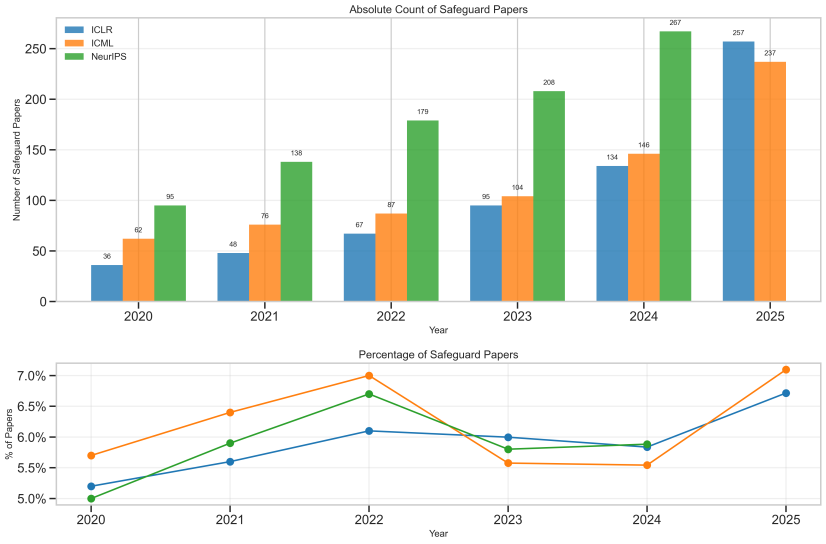

Figure 1: Percentage of accepted papers at ICML, ICLR and NeurIPS with key words indicating findings helpful for resisting AI-enabled authoritarianism. List of Keywords are in Table 3 in Appendix A.

and deployment of machine learning technology can, via the four technical properties discussed above, enhance authoritarian and degrade liberal-democratic features of political regimes, in particular through aggrandizement of executives and erosion of basic liberal rights. (For convenience, we summarily refer to this phenomenon as "AI-enabled authoritarianism.") We further argue for development of ML techniques that equip liberal-democratic institutions and citizens to resist AI-enabled authoritarianism.

By "AI systems" we mean technological systems engineered via statistical machine learning techniques that include both contemporary deep learning and "classical" ML algorithms. We reiterate that in this paper, "AI-enabled authoritarianism" is simply a way of referring to the ML-related violations of liberal norms we will discuss. We will not define "authoritarianism" intensionally or label existing political regimes as "authoritarian" or "liberal-democratic;" though we will speak of "more" and "less liberal-democratic" regimes, these descriptors are again meant only to conveniently invoke vague categories common in contemporary political discourse. The purpose of making such a distinction at all in this paper is to emphasize that AI-enabled authoritarianism exists even in those political regimes typically considered liberal-democratic. Of course, many of those regimes are also undergoing the broader political transformations that may lead to their reclassification by political scientists.

The scope of our survey is limited to two ways in which governments use AI technologies to surveil citizens and interfere with liberal democratic norms, specifically citizens' basic rights as generally defined in Articles 12 and 19 of the Universal Declaration of Human Rights. First, states collect information about citizens for purposes of targeting individuals who engage or might engage in behavior authorities aim to prevent. Second, states use surveillance technologies to control information flows in "cyber-space." Both kinds of surveillance can employ similar AI systems, and the first kind can rely on data collected from the internet. The cases differ in the structure of control: in the first case, authorities violate liberal norms by collecting and potentially acting on information *about* citizens, whereas in the second case, authorities violate liberal norms by controlling the flow of information *to* citizens.

Several important issues related to ML and the undermining of liberal democracy are outside the scope of this paper, though we do not claim that they are less important than the AI-enabled state surveillance we discuss. One such issue is the application of AI in foreign affairs: a state's use of AI technology can violate the putative liberal rights of foreigners not subject to that state's jurisdiction.

which are or previously were classed as (liberal) democracies according to the index [Varieties of Democracy (V-Dem) Institute, 2024].

Another such issue is the ML-occasioned concentration of power in private companies, described in, for example, Zuboff's *Age of Surveillance Capitalism*. In this paper, we discuss these topics only to the extent that they intersect very obviously with the direct exercise of state power over a state's own citizens in ways incompatible with mainstream conceptions of liberal democracy. A more comprehensive treatment of ML and liberal democracy would have to consider the topics much more broadly; international political economy, and much of domestic economics even in countries with more liberal-democratic political regimes, is illiberal and undemocratic.

After summaries of existing deployments of both kinds of AI-enhanced surveillance in several countries with diverse political regimes, we emphasize that practices like these occur in countries with political regimes often classified as liberal-democratic. We then introduce a three-pronged research agenda aimed at both ensuring that large-scale AI deployments meet basic technical standards of explainability and privacy preservation, and directly equipping citizens with adversarial tools to resist AI-enabled authoritarianism. We address alternative viewpoints.

2 Surveillance for intervention in "physical space"

Authorities may use AI-enabled surveillance to intervene physically and prevent or punish behavior that would be protected by standard negative liberal rights. Advances in ML allow governments to monitor citizen behavior at a much more granular level and even intervene preemptively.

Computer vision and biometric systems have transformed surveillance capabilities through real-time facial recognition and tracking, enabling authorities to monitor individual movements across entire cities. The system analyzes behavior patterns in public spaces, automatically flagging "suspicious" activities or unauthorized gatherings. The Chinese government's Sharp Eyes program represents the most comprehensive deployment of such a system, integrating over 200 million AI-enabled cameras into a national monitoring network for "100% coverage" [Feldstein, 2019]. Analogous systems are being deployed across a range of political regimes, though at varying scales and with different technological and legislative constraints. The Indian government's Central Monitoring System (CMS) provides telecommunications surveillance capabilities enhanced by AI analytics [Greenleaf, 2014]. Law enforcement agencies in multiple countries use the Clearview AI system for facial recognition from social media images [Rhinelander et al., 2024]. Further examples include the Russian government's System for Operative Investigative Activities (SORM) Soldatov and Borogan [2015], the Israeli military intelligence services' comprehensive surveillance network in Gaza and the West Bank [Ali, 2024], and Australia's digital identity systems and biometric border control systems [Department of Home Affairs, 2024].

Combining surveillance with predictive analytics enables preemptive, rather than reactive, intervention. In India, CMS enables the government to tap into communications at will, completely bypassing service providers. As a result, citizens have no way of knowing when the government has accessed their data [Internet Freedom Foundation, 2020]. Meanwhile, the Chinese Government has used its Integrated Joint Operations Platform (IJOP) to suppress dissent in Xinjiang. IJOP combines multiple data streams–from CCTV, WiFi, and police checkpoints–with predictive analytics to identify potential problems for authorities before they manifest. The platform flags "risky" individuals or groups based on behavioral patterns [Human Rights Watch, 2019, Watch, 2019]. These examples illustrate how AI-driven surveillance structures do not only increase the degree of centralized monitoring but also transform the operational logic from documenting transgressions to anticipating them. Such a process, could result in the "gradual disempowerment" of citizens whereby their default interests and preferences are disregarded and diminished by a power-optimizing state [Kulveit et al., 2025].

3 Surveillance for online information control

We turn next to the use of AI in online information ecosystems, where both classical ML and advances from the past few years have amplified authorities' abilities to manage information flows.

In Russia, state-sponsored cyber-espionage groups leverage automated monitoring within the SORM infrastructure to identify and suppress opposition content across social media platforms—often acting within minutes of content posting [Soldatov and Borogan, 2015, Polyakova and Meserole, 2019]. This represents a qualitative shift from previous censorship approaches, enabling automated, near-real-time narrative control at scale. Real-time content moderation systems powered by multimodal AI models

enable simultaneous analysis of text, audio, and visual content. Facebook's automated moderation and TikTok's recommendation algorithms demonstrate how these systems can shape information flows while maintaining a façade of algorithmic neutrality [Bradshaw et al., 2021].

Hungary's media-monitoring systems use AI-powered content analysis *de facto* to systematically suppress opposition voices while maintaining a veneer of [Howard and Bradshaw, 2020]. In Turkey, the 2020 social-media law requires large platforms to establish local offices and comply with government demands for content removal [19, 2025]. The Chinese government integrates AI-enabled information control with other surveillance infrastructure, exemplifying what is sometimes termed "networked authoritarianism" [Roth and Wang, 2019].

4 Surveillance Expansion in Relatively Liberal-Democratic Political Regimes

National security officials and owners of digital technology companies in countries with relatively liberal-democratic political regimes may want the ML research community and the broader public to believe that institutions including constitutional courts, data-protection authorities, and freedom-of-information laws insulate those societies from illiberal and anti-democratic practices of AI-assisted state surveillance. The justification for such belief is generally thinning as the international authoritarian political trend mentioned in the introduction includes the erosion of these institutions in some of the more liberal-democratic countries, a process some political scientists call "democratic backsliding" [Bermeo, 2016]. Moreover, the historical record of AI-assisted surveillance already tells a different story. The four technical affordances listed in the introduction—population-scale data ingestion **(A1)**, black-box inference **(A2)**, predictive automation **(A3)** and real-time execution speed **(A4)**—have repeatedly led to bypass of formal constraints. Table 1 summarizes patterns.

Technical feature	Deployment	Failure of liberal-democratic governance
Data ingestion	Bulk interception laws: UK *Investigatory Powers Act* (2016); French *Loi Renseignement* (2015)	Oversight bodies barred from inspecting raw datasets; retention periods up to 5 years
Black-box inference	Secret risk scoring: Dutch *SyRI* welfare fraud model; US TSA "Quiet Skies" traveller scoring	Parliamentarians learned of systems only after leaks; discriminatory features undetected for years
Predictive automation	PRE-CRIME pilots: Denmark "Gladsaxe" child-welfare system; Canada "Project Algorithmic Justice"	Administrative penalties issued on statistical suspicion, bypassing due-process hearings
Real-time speed	Live facial recognition at protests: UK Metropolitan Police; Australian states during COVID-19 lockdowns	Chilling effect on assembly; court review takes place months after deployments

Table 1: How the four technical features affect more liberal-democratic political regimes.

AI development and deployment for surveillance and control of citizens can be insensitive to formal liberal-democratic constraints for several interrelated reasons. One reason may be technical opacity. New technologies are sometimes so complicated that they overwhelm established oversight mechanisms. Additionally, AI systems based on deep learning, such systems have another layer of opacity because they are unexplainable even by the standards of computer science. In various cases, technical opacity such as that associated with black-box inference (**A2**) can either be a side effect of the best known techniques for accomplishing a certain goal (a goal whose social value may itself be disputable) or instead integral to a strategy to concentrate power. We discuss two other ways to understand AI-enhanced surveillance in more liberal-democratic countries.

4.1 The "security exception" to liberal democracy

In matters of "security," many states in nominally liberal-democratic political regimes already have a history of violating citizens' liberal rights and circumventing democratic processes [Lehr and Lehr, 2019]. This general phenomenon simply continues with AI technology providing new tools for

authorities. The national security exemption clause within the EU AI Act, for example, has left this regulatory space open to challenge and interpretation.

To see the continuity, consider the recent history of digital data collection for "security" purposes in more liberal-democratic countries. The Snowden archive exposed NSA bulk-collection programmes such as PRISM, implemented without congressional awareness [Miller and Walsh, 2016]. The European Court of Human Rights later condemned UK bulk interception [BBW, 2018], yet the Investigatory Powers Act re-legalized data collection under closed "technical capability notices."

In July 2024, the German newspaper Netzpolitik published a leaked document describing an EU proposal for mass surveillance, encryption backdoors, and enhanced cross-border cooperation measures [Netzpolitik.org, 2024]. The proposal, following a related one by the Swedish Presidency of the Council of Europe in the spring of that year, was drafted by the "High-Level Group (HLG) on access to data for effective law enforcement," a group that included senior officials from member states and the Commission, representatives of EU justice and home affairs agencies, and the EU Counter-Terrorism Coordinator, and was chaired by the Council Presidency and the Commission. Netzpolitik.org [2024]. The document, which has since been made public by the European Commission, begins by asserting the importance of preventing EU legislation from "interfering with national security," and contains 42 specific recommendations which, in essence, call for the revival of mass telecommunications surveillance via "data retention," the creation of state back-doors to encryption software, and increased cross-border cooperation, among other policies. Hundreds of academics and technical experts have criticized the irreconcilability of data-retention and client-side scanning with liberal rights to privacy and the presumption of innocence [Statewatch, 2023]. Furthermore, had the HLG working paper not been leaked to Netzpolitik, current EU protocol would not have made such documents available to the general public. Such episodes exemplify the *evidence* and *deployment* gaps: harms remain invisible until a whistleblower or litigant forces disclosure.

AI capabilities developed for military applications–which, we reiterate, have serious ramifications for liberal democracy that are outside our scope–migrate rapidly inward. Israel's 2024 Facial Recognition Bill and the National Cyber Directorate framework allow military-grade computer-vision pipelines to police domestic public spaces [fac, 2023, Directorate, 2021]. Similar technology transfer is underway in the United States, where a GAO report found federal agencies deploying facial-recognition services from military vendors without proper authorization [Wright, 2023].

4.2 Public–private surveillance assemblages

Besides the general imperviousness of "security" policy to democratic accountability and liberal norms, another factor in the expansion of illiberal and anti-democratic surveillance programs in more liberal-democratic countries is that commercial incentives reinforce state demand. Amazon Ring maintained more than 1,300 U.S. police partnerships before pausing its Law-Enforcement Request Portal in 2024. Freedom-of-information releases show Palantir's European public-sector contracts growing from tens of millions of euros in 2016 to hundreds of millions in 2024 (NHS FDP, Frontex, national MoDs) [Williams, 2021]. These arrangements cloak technical details behind trade-secret law, frustrating even well-resourced oversight committees. Globally, the surveillance-technology market is projected to exceed $300 billion by 2028, generating continuous pressure for deeper data access [Statista, 2023]. Meanwhile, the large platform companies are central to (state) surveillance expansion in the more liberal-democratic countries, as they both provide data and computing infrastructure for the specialized projects of smaller companies specializing in "security" technology, and collaborate directly in government surveillance programs, e.g., secretly allowing the NSA to access servers and collect user data under PRISM [Gellman and Poitras, 2013].

Implications for technical safeguards

Liberal-democratic institutions provide at best partial resistance where "security," commercial interests, or both are at play. The safeguards proposed in Section 5 directly target those voids:

- **Scalable differential privacy** raises the statistical cost of bulk retention, limiting **A1**.
- **Certified explanations** give courts and civil society leverage against black-box scoring (**A2**) and predictive automation (**A3**).

- **Adversarial user tooling** restores some agency when real-time systems (**A4**) are fielded without consent.

Absent such technical reinforcement, the formal checks of liberal democracy will continue to erode under pressure from both state security demands and commercial surveillance incentives.

5 Explainable-by-Design, Privacy-by-Default: A Three-Pronged ML Research Agenda

Sections 2–4 uncovered four persistent *gaps*: *evidence*, *capability*, *deployment* and *asymmetry*. Closing them requires technical work in three complementary areas. Table 2 situates each research thrust within those gaps, names concrete ML problems and flags the main obstacles and risks.

Research thrust	Gap(s)	Concrete ML problems	Why unsolved	Risks / limits
Scalable privacy preservation	Capability, Deployment	DP training for 10B-parameter models; cryptographically verifiable audit logs; Byzantine-resilient FL on non-IID data	DP drops accuracy; audit primitives brittle; existing FL theory assumes IID	Utility loss; adaptive adversaries
Formal interpretability & causal explanation	Capability, Evidence	Counterfactual explainers for multimodal transformers; certified concept attribution; online detection of feature inversion	XAI does not scale; no causal guarantees; high compute cost	Sensitive-feature leakage; explanation gaming
Adversarial user tooling	Asymmetry, Deployment	Real-time face scrambling on mobile; gradient-free traffic morphing; wearable perturbation learning	Latency budgets; breaks under model updates; UX friction	Arms race; platform bans

Table 2: Three technical thrusts and how they close the gaps identified earlier.

5.1 Thrust 1 – Scalable Privacy Preservation

How it counters AI-enabled authoritarianism. Sections 2 and 3 showed that mass data retention fuels face-recognition grids and predictive policing. Strong, *provable* privacy guarantees force aggressors to pay a statistical cost for each additional data record, thereby capping the surveillance payoff. Public audit logs further enable NGOs and journalists to verify compliance, closing the *evidence* gap.

Shortcomings & caveats. Differential-privacy budgets remain difficult to communicate to non-experts; tight budgets raise error rates that may undermine adoption; adversaries can still combine leaked aggregates with auxiliary data. Encryption or DP alone does not stop coercive endpoints such as mandatory camera installation.

Priority problems. (a) DP-optimisation schedules for billion-parameter transformers with $< 3\%$ accuracy loss; (b) attested vector-commitment audit logs that append a zero-knowledge proof to each inference; (c) Byzantine-resilient federated learning that tolerates malicious sensors without abandoning privacy budgets.

5.2 Thrust 2 – Formal Interpretability and Causal Explanation

How it counters AI-enabled authoritarianism. Black-box inference (feature *ii*) strips courts and legislators of epistemic leverage. Certified explanations allow external actors to *demonstrate* discriminatory feature reliance and to contest automated decisions, curbing the executive power aggrandisement documented in Section 4.

Shortcomings & caveats. No explanation is fully neutral: revealing internal logic can leak sensitive attributes and enable adversarial reverse-engineering. Causal certificates depend on untestable

assumptions about the data-generating process; bad faith actors may disclose only favourable slices of an explanation.

Priority problems. (a) Causal surrogate models for vision–language transformers with bounded counterfactual risk; (b) PAC-style concept certificates that guarantee protected attributes influence the logit by no more than a user-set threshold; (c) lightweight streaming rationales for real-time moderation, delivered under 50 ms.

5.3 Thrust 3 – Adversarial User Tooling

How it counters AI-enabled authoritarianism. Because authoritarian actors already enjoy a data and resource surplus (*asymmetry gap*), even marginal defensive tools can change outcomes for individual activists. Real-time perturbations—visual, acoustic or traffic-based— restore some agency while policy lags behind technology.

Shortcomings & caveats. Cloaking shifts responsibility onto individuals rather than institutions; an active arms race can normalise heavier surveillance; usability barriers mean disadvantaged groups may benefit least. Platforms might ban perturbation tools as "malware", resurrecting dependency on corporate goodwill.

Priority problems. (a) Perceptually consistent face-ID cloaking robust to future model updates; (b) gradient-free padding schemes that obfuscate packet sequences under tight mobile bandwidth; (c) human-in-the-loop perturbation learning that trades off attack strength against social acceptability.

6 Alternative viewpoints

Building technical safeguards is not a universally accepted solution. Below we summarize four common objections and offer brief replies, highlighting where our proposal is complementary to, rather than in conflict with, broader policy and social strategies.

Objection 1: *"Policy, not technology, is the real bottleneck."*

Reply. Sections 4 and 5 acknowledge that formal privacy law and oversight are indispensable for liberal democracy. However, many jurisdictions lack the forensic capacity to verify whether black-box systems comply with policy once it is enacted. Scalable differential privacy and certified explanations provide an empirical foothold for regulators and courts, making mandates enforceable.

Objection 2: *"Technical safeguards legitimize applications that should simply be banned."*

Reply. We agree that certain applications may warrant outright bans. In those cases, the adversarial tools in Thrust 3 remain important both before the bans are instituted and afterward when they may not be enforced for reasons outlined in Section 4. The other two thrusts in our agenda target cases where society has *not yet* reached consensus or where outright prohibition is, at least for the moment, politically infeasible. Here, strong privacy and interpretability can reduce harm while normative debates and political struggles continue.

Objection 3: *"An adversarial arms race is futile; the corporate state will always win."*

Reply. Even partial degradation of surveillance accuracy can raise the cost of repression or create evidentiary doubt in court. Historical examples—from PGP to Tor—show that inexpensive defensive tools can significantly shift power toward citizens. Our research agenda seeks to make such tools available. A longer-term goal would be for masses of citizens to effect political-economic changes that would reduce the need for adversarial competition in ML by reducing concentrations of power and reducing incentives for illiberal applications of technology.

Objection 4: *"Explanation techniques leak sensitive features or can be gamed."*

Reply. Indeed, techniques such as naïve post-hoc saliency maps can leak private data and are vulnerable to manipulation. That is why Thrust 2 prioritises *formal* interpretability with statistical guarantees and parallel red-teaming. The aim is not to publish raw model internals, but to provide bounded, verifiable evidence that protected attributes do not dominate decisions.

On the other hand, there will be objections that sound similar to this one but are actually about the protection of corporations' intellectual property. Recall from Section 4 that the current intellectual property regime contributes to AI-enabled authoritarianism in the more liberal-democratic countries. This is where the spirit of Objections 1 and 2 becomes quite relevant. ML researchers should engage the broader public in discussion about algorithmic transparency, rather than allowing narratives to be set by business and policy elites.

Objection 5: *"Decentralization makes centralized safeguards unnecessary."*

Reply. We agree that decentralized model ownership presents a powerful structural alternative to centralized surveillance infrastructure. Running models locally on personal devices rather than remote servers makes privacy the default by design and constrains the reach of institutional actors. As model compression and inference-time optimization advance, this path becomes increasingly feasible: capable open-source models are now runnable on consumer hardware, and further progress in distillation, quantization, and low-rank adaptation will widen access. Notably, these are areas of capabilities research that directly advance democratic values by enabling broad-based control over ML systems rather than exclusive control by corporate or state actors.

However, decentralization does not obviate the need for centralized safeguards. Most deployed ML systems today remain cloud-based, centrally managed, and opaque. Meanwhile, decentralized deployments introduce new risks, such as uneven access to protective tools, lack of coordination mechanisms, and vulnerability to local coercion. Our agenda is therefore complementary: technical progress on decentralization is a critical line of resistance, but formal guarantees for centralized systems remain an urgent priority. The field should pursue both.

Synthesis. Technical safeguards are neither a silver bullet nor a distraction from governance; they are a necessary layer that empowers courts, journalists, civil-society technologists and, ultimately, activists and the public to hold powerful actors accountable [Watson et al., 2024]. Effective democratic defense therefore requires a *policy–technology complementarity*, not a false choice between the two.

7 Conclusion

Artificial intelligence built on statistical learning has already re-wired surveillance and information control. Its four core features—population-scale data ingestion, black-box inference, predictive automation and human-out-of-the-loop speed—tilt power toward executives, undermine liberal rights, and compress the public sphere. Our case studies (Sections 2–4) illustrate how this dynamic operates both in overtly authoritarian regimes and in nominally liberal-democratic states that rely on secrecy, emergency framing and cooperation with private technology companies.

We therefore advance a three-pronged research agenda (Section 5) that re-prioritises mainstream ML effort toward *privacy preservation*, *formal interpretability* and *adversarial user tooling*. Each thrust is mapped to the evidence, capability, deployment and asymmetry gaps that currently enable ML-powered authoritarianism, and each comes with explicit caveats and red-teaming requirements.

Limitations. Technical safeguards alone cannot abolish illiberal practices; determined authorities may still coerce data at collection points, outlaw perturbation tools or ignore audit evidence. But without verifiable privacy, certified explanations and accessible defensive tooling, democratic institutions and civil society lack the factual leverage and autonomy needed to contest abuse. Our agenda is thus simply a prerequisite for durable liberal-democratic defense.

Call to action. The NeurIPS, ICML and ICLR conferences could reserve scored tracks for safeguard breakthroughs. Funding agencies should create dedicated grant lines for defensive ML to match current capability-oriented funding. In the event of such institutional changes to research incentives, ML researchers should vigilantly ensure that those arrangements do not simply become vehicles for the waging of international economic and military competition under the pretext of "anti-authoritarianism." Researchers should benchmark progress on datasets such as DP-LLAMA, CAUSAL-IMAGENET and CLOAK-CHALLENGE. Even small shifts in community effort toward the

goals we have outlined–while researchers also keep in mind the legitimate concerns represented in Section 6 regarding this kind of research agenda–could generate substantial social progress.

Explainable-by-Design, Privacy-by-Default should no longer be optional add-ons; they must become the default expectation for every large-scale ML system. Only then can policy, oversight, and civic action stand a fighting chance against the accelerating tide of AI-enabled authoritarianism.

References

The state of surveillance in 2018. Technical report, Big Brother Watch, London, UK, 2018. URL https://bigbrotherwatch.org.uk. Accessed: 2025-05-23.

Israeli government backs facial recognition legislation. *Legal Affairs Monitor*, September 2023. Analysis of facial recognition legislation.

Article 19. Regulation of social media platforms in turkey, 2025. URL https://www.article19.org/resources/regulation-of-social-media-platforms-in-turkey-internet-law/. Accessed: 2025-03-13.

Nijmeh Ali. Israeli online surveillance regime. *Resisting Domination in Palestine: Mechanisms and Techniques of Control, Coloniality and Settler Colonialism*, page 51, 2024.

Nancy Bermeo. On democratic backsliding. *Journal of democracy*, 27(1):5–19, 2016.

Samantha Bradshaw, Hannah Bailey, and Philip N. Howard. Industrialized disinformation: 2020 global inventory of organized social media manipulation. Technical report, Oxford Internet Institute, 2021. URL https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2021/02/Industrialized-Disinformation-Report-2021.pdf.

Department of Home Affairs. Digital identity and biometric capability program: Annual report, 2024. Details Australia's implementation of AI-enabled border control systems.

Israel National Cyber Directorate. Facial recognition in public places. Technical report, July 2021. URL https://www.gov.il/en/pages/face_recognition. Discusses advancements in facial recognition technology and its broader applications in public security.

Steven Feldstein. The road to digital unfreedom: How artificial intelligence is reshaping repression. *Journal of Democracy*, 30(1):40–52, 2019.

Barton Gellman and Laura Poitras. U.s., british intelligence mining data from nine u.s. internet companies in broad secret program, June 2013. URL https://www.washingtonpost.com/investigations/us-intelligence-mining-data-from-nine-us-internet-companies-in-broad-secret-program/2013/06/06/3a0c0da8-cebf-11e2-8845-d970ccb04497_story.html. Accessed: 2025-05-23.

Graham Greenleaf. *Asian Data Privacy Laws: Trade and Human Rights Perspectives*. Oxford University Press, 2014.

Philip N Howard and Samantha Bradshaw. The global disinformation order: 2020 global inventory of organised social media manipulation. *Computational Propaganda Research Project*, 3:1–45, 2020.

Human Rights Watch. China's algorithms of repression: Reverse engineering a xinjiang police mass surveillance app, 2019.

Internet Freedom Foundation. Watch the Watchmen Series Part 2: The Centralised Monitoring System, 2020. URL https://internetfreedom.in/watch-the-watchmen-series-part-2-the-centralised-monitoring-system/. Accessed: 23 May 2025.

Jan Kulveit, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, and David Duvenaud. Gradual disempowerment: Systemic existential risks from incremental ai development, 2025. URL https://arxiv.org/abs/2501.16946.

Peter Lehr and Peter Lehr. Undemocratic means: The rise of the surveillance state. *Counter-Terrorism Technologies: A Critical Assessment*, pages 169–179, 2019.

Seumas Miller and Patrick Walsh. The nsa leaks, edward snowden, and the ethics and accountability of intelligence collection. In *Ethics and the Future of Spying*, pages 193–204. Routledge, 2016.

Netzpolitik.org. Going dark: Eu-expertengruppe fordert hindertüren und mehr überwachung, 2024. URL https://netzpolitik.org/2024/going-dark-eu-expertengruppe-fordert-hintertueren-und-mehr-ueberwachung/. Accessed: 2025-01-12.

Alina Polyakova and Chris Meserole. Exporting digital authoritarianism: The russian and chinese models. Technical report, Brookings Institution, 2019. URL https://www.brookings.edu/research/exporting-digital-authoritarianism-the-russian-and-chinese-models/.

Jason Rhinelander, Claudia De Fuentes, and Cynthia O'Driscoll. Clearview ai: ethics and artificial intelligence technology. In *Cases on Entrepreneurship and Innovation*, pages 237–246. Edward Elgar Publishing, 2024.

Kenneth Roth and Maya Wang. Data leviathan: China's burgeoning surveillance state. *The New York Review of Books*, August 2019. URL https://www.nybooks.com/online/2019/08/16/data-leviathan-chinas-burgeoning-surveillance-state/.

Andrei Soldatov and Irina Borogan. *The Red Web: The Struggle Between Russia's Digital Dictators and the New Online Revolutionaries*. PublicAffairs, 2015.

Statewatch. Statement to eu countries: Do not agree to mass surveillance proposal, warn ngos, September 2023. URL https://www.statewatch.org/news/2023/september/statement-to-eu-countries-do-not-agree-to-mass-surveillance-proposal-warn-ngos/. Accessed: 2025-01-12.

Statista. Surveillance technology market size worldwide 2027, 2023. URL https://www.statista.com/statistics/1251839/surveillance-technology-market-global/. Accessed: 2025-05-23.

Varieties of Democracy (V-Dem) Institute. Democracy report 2024: Autocratization turns viral, 2024. URL https://www.v-dem.net. Accessed: 2025-01-19.

Human Rights Watch. China's algorithms of repression: Reverse engineering a xinjiang police mass surveillance app. *Human Rights Watch Report*, 2019. URL https://www.hrw.org.

David S. Watson, Jakob Mökander, and Luciano Floridi. Competing narratives in ai ethics: a defense of sociotechnical pragmatism. *AI & Society*, December 2024. doi: 10.1007/s00146-024-02128-2. URL https://doi.org/10.1007/s00146-024-02128-2. Accessed: 2025-05-23.

Martin Williams. 'spy tech' firm palantir made £22m profit after nhs data deal, August 2021. URL https://www.opendemocracy.net/en/dark-money-investigations/spy-tech-firm-palantir-made-22m-profit-after-nhs-data-deal/. openDemocracy.

Candice N. Wright. Facial recognition technology: Federal agencies' use and related privacy protections. Technical report, U.S. Government Accountability Office, 2023. Statement of Candice N. Wright, Director, Science, Technology Assessment, and Analytics.

Appendix

Keyword list

Category	Keywords Used to extract data from OpenrReview API
Privacy & Differential Privacy	differential privacy; local differential privacy; LDP; Rényi differential privacy; privacy accountant; privacy amplification; membership inference; model inversion; privacy leakage; machine unlearning; data deletion; right to be forgotten; secure multiparty; SMPC; multi-party computation; zero-knowledge; ZKP; trusted execution; TEE; SGX; homomorphic encryption; FHE; CKKS; Paillier
Federated & Robust FL Security	federated learning; FedA; FedD; FedF; FedN; FedO; FedP; FedProx; FedDyn; FedNova; FedDF; secure aggregation; SecAgg; robust aggregation; median aggregation; trimmed mean; Krum; Bulyan; model poisoning; backdoor attack; backdoor defence; Sybil
Interpretability & Explainability	interpretability; interpretable; mechanistic interpretability; circuits analysis; attention head; neuron activation; representation probing; latent probing; explainable; explainability; XAI; SHAP; LIME; integrated gradients; GradCAM; saliency map; feature attribution; counterfactual explanation; contrastive explanation; concept bottleneck; concept activation; algorithmic auditing; model auditing; responsible AI
Citizen-side Privacy Tools	adversarial patch; invisibility cloak; advcloak; stylecloak; face obfuscation

Table 3: Keyword list used to identify *safeguard-oriented* papers.