

AI-Enabled Authoritarian Practices in Formal Democracies

Keir Reid*, Isaac Friend*, Fazl Barez

Oxford Martin AI Governance Initiative
University of Oxford

isaac.friend@oxfordmartin.ox.ac.uk, fazl@robots.ox.ac.uk

Abstract

Western political scientists' treatment of digital authoritarianism has mostly been anchored in the concept of "authoritarian regime." Constantly framing discussions in terms of regime type deflects attention from authoritarian phenomena in Western formal democracies. Moreover, despite accounting for some advances in machine learning, the literature often falls behind rapid technological innovation. This paper applies a "practice"-based definition of authoritarianism to identify and analyze AI-enabled authoritarianism in democratic systems. We trace historical continuity in authoritarian practices in the national security apparatus of democratic states, while also describing new modes of authoritarianism made possible by deep learning. Characteristics of AI that surveillance and privacy scholars have argued occasion new kinds of social power are here shown to be specifically authoritarian affordances. These affordances can be used by power-wielders within a democratic state to disable the voice and information access not only of the public at large, but of intra-government actors tasked with holding those power-wielders to account.

Introduction

For more than a decade, political scientists have commented on the relationship between digital technologies and authoritarian politics, often documenting how digital technologies have expedited governments' practices of surveillance and repression of citizens in what are called authoritarian political "regimes" (Dragu and Lupu 2021; Erixon and Lee-Makiyama 2011; Frantz, Kendall-Taylor, and Wright 2020; Lamensch 2021; Polyakova and Meserole 2019). Writers refer to the general phenomenon as "digital repression" or "digital authoritarianism."

The most obvious problem with the mainstream concept of digital authoritarianism is that "authoritarianism" is considered a property attributable to political regimes, and is further defined, explicitly or implicitly, in such a way that the regimes of Western countries are not labeled "authoritarian." Thus, the mainstream concept exempts similar practices when they occur in political regimes that Western political scientists and commentators choose not to classify as authoritarian.

*These authors contributed equally.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

In regime-focused literature on digital authoritarianism, it is often difficult to discern what role the "regime type" actually plays in the analysis. Some of this literature can therefore be understood as part of a pattern of Western academic and policy elites superficially associating human rights problems in certain countries with the absence in the countries' political regimes of formal characteristics such as elections that are said to distinguish (liberal) democracy from authoritarianism. This pattern may function either to divert attention from illiberal and undemocratic goings-on within the formally liberal-democratic political regimes in those countries—and other ways in which the societies more broadly are illiberal and undemocratic—or at least to suggest, without justification, that illiberal and undemocratic tendencies should be considered essential to politics in some countries but not others. In the case of digital authoritarianism, the analytical selectivity may obscure structural parallels among formally diverse settings and limit our ability to understand the impact of emerging technology across the spectrum of existing formal political regime types.

Some authors depart from the focus on authoritarian regimes either by intentionally de-emphasizing the term "authoritarianism" in favor of "repression" (Feldstein 2021) or by studying authoritarian *practices* rather than regimes (Glasius and Michaelsen 2018; Hintz and Milan 2018; Maerz 2026). (We take the latter route in this paper.) Whether regime-focused or not, however, literature on digital authoritarianism has not been particularly focused on the role of AI and is often more generally concerned with all manner of digital data collection and processing by authorities.

Meanwhile, another corpus of scholarship, represented most prominently by books such as Shoshana Zuboff's *Age of Surveillance Capitalism* (2019) and Kate Crawford's *Atlas of AI* (2021), directly addresses (among other issues) the anti-democratic politics of AI in countries with formally democratic political regimes, especially the United States.¹

¹In this paper, "formal democracy" means "formally liberal-democratic political regime." There is potential for initial confusion as we will distinguish authoritarianism from illiberalism and concentrate on the former, even while eliding the distinction between democracy and liberalism by using "democracy" as shorthand. The point is that we are analyzing authoritarian rather than illiberal practices, but the countries in which the practices occur have liberal-democratic political regimes. When a regime is for-

The type of social scientific thought displayed in these works correctly locates AI politics elsewhere than in the form of a national political regime. Specifically, it studies the embedding of AI in systems of social control in service of wealth extraction and accumulation. It attends to ways in which the techniques of actual, statistical learning-based AI feed into or affect the “logic” of social control, or occasion new kinds of power, rather than simply how increases in “capabilities” of digital technology give the owners of those systems “more power” (as in discussions of AI-generated “deepfakes” influencing elections). In situating states within larger power systems often dominated by corporations, however, the political-economic critiques have tended not to concentrate on state authoritarianism *per se*.

Hence, the study of digital technology and politics could be advanced by a discussion that applies insights from various social sciences to understand how today’s deep learning-based AI contributes to state authoritarianism, positively and narrowly defined, in contexts where formal democracy, on the whole, appears stable to regime-focused political scientists. That is the contribution of the present paper, whose examples are drawn from the United States and the United Kingdom. We organize our study on the basis of the political scientist Marlies Glasius’s (2023) conception of authoritarian practices as accountability sabotage, and two subsidiary conceptual frameworks due to other authors: the national security “double black box” (Deeks 2025) and the conformity theory of chilling effects (Penney 2025). We analyze, with ideas from these frameworks, paradigmatic examples of authoritarian national security practices from the Cold War and the Global War on Terror, commenting on how pre-AI technological affordances shaped authoritarianism in formal democracies. Finally, we arrive at the present AI era, in which deep neural networks’ predictive accuracy useful in multiple task domains—particularly through language modeling—and their algorithmic opacity function as authoritarian affordances. An under-studied issue even in the few prior non-regime-based discussions of authoritarian practices in formal democracies, and one we especially emphasize, is that of authoritarian practices directed by a configuration of national security actors at another configuration of government actors (even within the same branch of government) that would attempt to represent the public by holding the former actors to account.²

Authoritarian Practices

We follow Marlies Glasius in defining an authoritarian practice as “*a pattern of actions, embedded in an organized context, sabotaging accountability to people over whom a configuration of actors exerts a degree of control, or their rep-*

mentatives, by disabling their voice and disabling their access to information” (Glasius 2023, p. 22). Combining different precedents noted by Glasius, we call the configuration of actors exerting control the *power-wielder* (Grant and Keohane 2005), and the people at the other end of the accountability relationship the *forum* (Bovens 2007). For cases of authoritarian practices directed immediately at representatives of the forum, we introduce the term *checker* to refer to the representatives.

As Glasius notes, the definition solves problems with notions of authoritarianism popular in political science. It is positive, rather than defining authoritarianism in terms of the absence or opposite of democracy. And it avoids associating authoritarianism primarily with the conduct of elections. Glasius excavates from the history of political thought what was considered important about democracy and about free and fair elections in particular, and what was said to be undermined by authoritarianism: accountability of rulers to those governed.

The definition is usefully restrictive in two ways. First, it distinguishes an authoritarian practice from a practice that simply takes place without accountability. Authoritarianism is in the sabotage and not the absence of accountability. “*Accountability sabotage coexists with, and pre-empts or responds to, demands for accountability*” (Glasius 2023, p. 21). Second, the definition specifies the two relevant types of accountability sabotage practices. Both of these restrictions indicate the minimal notion of “accountability” being used: accountability of a configuration of actors to a forum over which the actors wield power means roughly that the forum has the information necessary to judge the power-wielder’s conduct and can voice questions and judgments, not that the forum influences the power-wielder’s actions. Our term *checker* does suggest some influence on the power-wielder’s actions, and often applies to situations—such as congressional oversight of executive activities—in which the checker is formally responsible for exercising such influence. In such cases, the issue with respect to authoritarianism is not simply whether the checker has enough power to counteract the power-wielder, but whether the power-wielder interferes with the checker by disabling voice or information access.³

The restricted compass of the definition of authoritarianism allows for a further conceptual distinction that Glasius also endorses: between authoritarian practices and illiberal ones. Illiberal practices are those that infringe on individuals’ autonomy and dignity. They are only authoritarian insofar as they constitute accountability sabotage. On the other hand, authoritarian practices may or may not be illiberal, depending on whether the practices of accountability sabotage also directly infringe on individual autonomy and dignity.

³Sometimes authoritarianism vis-à-vis a checker may be more complex than interference at the time of the checker’s attempts to fulfil its function in a particular case. For example, if the power-wielder predates the checker, the power-wielder may have influenced the design and creation of the checker, and may at that time have engaged in authoritarian practices vis-à-vis the forum, which would have preferred more open dialogue about the design of the checker.

²We do not deal with practices specific to the political transformation taking place in the United States under the second Trump administration, as mainstream political scientists are describing these practices in regime terms (Carpenter 2026).

The conceptual distinction depends on considering only the immediate consequences of a practice, since ultimately the status of individual autonomy and dignity in a social context and the stability of accountability mechanisms in that context are codependent. One example of an illiberal practice that is not necessarily authoritarian is arbitrary surveillance. We will, however, consider many cases of surveillance that is used for authoritarian purposes, such as intervention to prevent expression and thereby disable voice. Surveillance practices are often authoritarian in their secrecy, which disables parties' access to information.

Defining AI-Enabled Authoritarianism

By AI-enabled authoritarianism we mean authoritarian practices in which the power-wielder uses information technology tools engineered with statistical machine learning techniques. We especially have in mind AI systems of the present deep learning era,⁴ whose architecture produces both increases in predictive accuracy and a certain kind of opacity: the inability to recover human-readable rationales for an outcome, because deep neural networks apply successive nonlinear transformations across millions of numerical parameters whose relationship to any specific input cannot be traced through inspection alone.

The concept is meant to capture authoritarian practices that rely rather directly on AI (understood as a set of tools), not all of the ways in which AI may change society and lead to future authoritarianism. Complementary to the notion of AI-enabled authoritarianism is what might be called "AI-driven" authoritarianism, where the term "AI" switches referent from a set of technologies to a social agenda. That is, power-wielders' goals to develop and use certain kinds of AI products drive the power-wielders to engage in authoritarian practices in service of those goals. The practices themselves may or may not rely on existing AI technology, i.e., may or may not be AI-enabled.

State-Corporate Authoritarianism

The definition of authoritarianism excludes from our study several forms of illiberal and undemocratic arrangements. For example, plutocracy, a main mode of national governance in many countries today, and one requiring special attention in connection with AI, is not authoritarian in itself, though it may be accompanied by authoritarian practices in anticipation of or response to citizens' opposition. On the other hand, the definition of authoritarianism is usefully expansive in that it does not specify the national government as the only source of authoritarian practices. Thus one can consider authoritarian practices in sub-national as well as national governments. This possibility is relevant, for example, to AI use by local police forces. Our main mention of AI in policing, however, arises in discussion of the agendas of

power-wielders in national government. We leave aside the topic of authoritarian practices in transnational governance arrangements.

Glasius's concept also allows for corporations as power-wielders engaged in authoritarian practices. Because of the distinction between authoritarianism and illiberalism and the minimal notion of "accountability" used in the idea of authoritarianism as accountability sabotage, corporate authoritarianism in this conception is not the same as the authoritarianism attributed to employers in anti-capitalist social thought—the employment relation's alleged undermining of employees' autonomy and dignity would be called illiberal under our definitions. In our sense, a corporation, or a configuration of actors wielding the corporation's power, may engage in authoritarian practices vis-à-vis the corporation's employees or customers, a government, or a broader public, and the entity "receiving" the authoritarian practices may have either the role of the forum or the role of the checker.

This paper is concerned with authoritarian practices in which a substantial amount of the power being wielded is state power, and the forum consists of citizens of the state. Corporations, however, are often involved in AI-enabled authoritarian practices as well as in prior digital authoritarian practices. It often makes sense to consider state and corporate interests convergent, and state and corporate power as mutually constitutive or operating in tandem. In such cases one can speak of state-corporate authoritarianism. Our treatment of the subject does not adjudicate between, and is compatible with, various theses about the causal priority of state versus corporate interests in state-corporate authoritarianism. Many contemporary social issues studied by "AI ethics and society" researchers can be understood in terms of state-corporate authoritarianism that is both AI-enabled and AI-driven.

Any serious mapping of the political economy of state-corporate AI-enabled authoritarianism would assign a more central place to the big tech owners of cloud infrastructure and digital platforms than to smaller specialist vendors of particular software products. We do not address these relationships at all in the present paper, viewing companies just as the suppliers of AI products to states.

The Double Black Box and Chilling Effects

While Glasius's concept of authoritarian practices as accountability sabotage organizes this paper, we rely on and synthesize parts of two other recently proposed conceptual frameworks to identify and interpret evidence of AI-enabled authoritarian practices. First is Ashley S. Deeks's framework of the "double black box" for considering the effects of AI-occasioned opacity on accountability in national security (Deeks 2025). Focusing on the US national security state—but providing insights that will apply to the security apparatus in many formally democratic settings—Deeks identifies a layering of two forms of opacity that creates a nearly impenetrable barrier to democratic oversight. The first component of the double black box is the traditional apparatus of national security secrecy, including classification systems, the state secrets privilege, executive privilege, and numerous less legally codified practices shielding se-

⁴That is, in exploring AI-enabled authoritarianism we do not seek to define AI abstractly, and much of what we say would not apply to rule-based systems (Haugeland 1985), since they are not at bottom statistical predictors and since they yield human-readable rationales for outputs in a way that deep learning-based systems cannot.

curity operations from public view, congressional (or parliamentary) oversight, and judicial review. Deeks notes that this secrecy often extends beyond the (already heavily influenced by power-wielders) bounds of “legitimate” secrecy to conceal illegality or incompetence and limit embarrassment, echoing Glasius’s indication of “illegitimate secrets” as a form of accountability sabotage (Glasius 2018, p. 532). While some of Deeks’s examples evince active accountability sabotage and Deeks mentions competing interests between national security officials and the public, she tends toward viewing opacity as a mere accountability void rather than as a tool of accountability sabotage, and accordingly does not invoke any notion of authoritarianism. Already one can see, however, that not only is national security opacity a strong candidate for an instance of authoritarian accountability sabotage, but the choice to analyze practices rather than political regimes is paying off, because in cases of national security accountability sabotage, the power-wielder and the checker whose voice or information access is being disabled are both government entities. Furthermore, even talk of “authoritarians” aggrandizing the executive branch by sabotaging accountability mechanisms involving the legislative and judicial branches is insufficient to understand the authoritarian qualities of the national security black box, because even checkers within the executive branch itself can experience the outside of the black box (Deeks 2025, p. 50).

The national security black box has existed in many technological eras. The novelty is the addition of an “algorithmic black box” of technical opacity due to AI. Layers of algorithmic opacity include—in addition to “the inability to know that an actor is employing an algorithm in the first place (which [Deeks] treats as part of the ‘national security black box’ problem)”:

the inability to obtain information about the type of data on which a programmer trained an algorithm; the inability to know the algorithm’s code; the inability to learn how the algorithm weighed values to reach an outcome, due to the highly complex nature of neural networks; the inability to understand how the algorithm learns and evolves; and the inability to know how accurate the algorithm’s recommendations are (Deeks 2025, p. 75).⁵

Deeks’ commentary focuses on the ways in which the resultant double black box threatens public law values; we will explore the double black box’s role in authoritarian accountability sabotage specifically. The second major interpretive frame we use is Jonathon W. Penney’s theory of chilling effects as conformity and compliance effects (Penney 2025). Penney contends that legal scholars’ conventional understandings of chilling effects are too narrow and inadequately grounded in behavioral science and social theory. Drawing on decades of empirical research in social psychology, he proposes to understand chilling effects—“where people are deterred or discouraged from speaking or acting freely” (Penney 2025, p. 1)—as inducements to con-

formity and compliance with norms, driven by four socio-psychological factors: observation, uncertainty, personalization, and authority. That is, people’s behavior is chilled when they sense that they may be observed, when they are uncertain about their situation, when they are threatened in a way that feels personally targeted, or when they are in a scenario involving entities with power and authority. Penney, a scholar of technology and society, emphasizes chilling effects created by governments’ and corporations’ use of digital technology for surveillance. Although he does not distinguish illiberal from authoritarian practices, we draw on his theory and evidence to highlight the production of chilling effects as a kind of authoritarian practice. Glasius and Michaelson, citing Penney’s earlier work (Penney 2016) among others, classify the chilling effects of illiberal surveillance as possible downstream consequences for democratic accountability, rather than as elements of authoritarian practices (Glasius and Michaelson 2018, p. 3802). In this case our result from applying Glasius’s concept differs from Glasius and Michaelson’s. Although in some instances of surveillance there is no evidence that the actors responsible intend the surveillance itself to disable voice by chilling expression—especially if the surveillance is kept secret from those surveilled—Penney’s work demonstrates that chilling effects produced by observation are actively engineered by powerful actors.⁶

Our main extension of Penney’s work is in considering weaponized chilling effects on checkers and not just on the public. In the context of the national security double black box, we can study, and construct likely narratives about, how power-wielders use information technology (and AI in particular) to chill checkers within government from accessing or disseminating information in ways the power-wielders do not want. We can recognize a different authoritarian function of the national security double black box: in addition to directly disabling checkers’ access to information, it creates uncertainty for checkers and thereby chills their behavior. Applying Penney’s insight that chilling effects are productive and not only repressive, we can also understand how power-wielders within the national security black box may chill potential checkers into cooperation with the power-wielders. Both the repressive and productive chilling of checkers raise a point that Penney makes in other contexts (Penney 2025, p. 103): a subject of a chilling effect may face competing sets of norms. Checkers face the norms of the power-wielder, the norms of the society at large, and norms encoded in procedural rules they are supposed to follow. A power-wielder’s ability to chill a checker into conformity with the power-wielder’s norms depends on the force with which the checker experiences the competing norms.

⁵Many researching neural network interpretability and explainability might be inclined to omit the first two list items from the “technical opacity” group.

⁶The caveat is that in Penney’s most complete example, that of Texas’s Senate Bill 8 to prevent abortion, the chilling effect is not directly on expression or on the act of accessing information relevant to judging the power-wielder. But Penney also convincingly argues that AI-enabled surveillance is and will be used to chill political expression and information access.

Pre-AI Authoritarian Practices in Formal Democracies

We will establish a basis for comparison of AI-enabled authoritarianism by discussing historical examples of state authoritarian practices in formal democracies. Throughout the twentieth and twenty-first centuries, accountability sabotage originating within states' national security apparatus has recurred in these regimes. Labeling adversaries as existential threats and abusing channels of state secrecy, configurations of national security and political elites have repeatedly disabled the voice and information access of both ordinary citizens and intra-government checkers.

The Cold War

The FBI's COINTELPRO, initiated in 1956, is a paradigmatic example of authoritarian practices in a formal democracy. A power-wielding configuration inside the Bureau used covert and frequently illegal means to disable the political voice of US citizens. These operations, which included physical intervention against, and disinformation about, domestic civic and political organizations, were rationalized as efforts to counter foreign interference and prevent violence, but the 1976 Church Committee Report found that the program had the unstated purpose of "maintaining the existing social order" (Select Committee to Study Governmental Operations with Respect to Intelligence Activities 1976, p. 7). The concealment of COINTELPRO from checkers is also well documented. The Church Committee Report found that FBI officials routinely omitted details of the covert activities when sharing intelligence with other parts of the executive branch, and that even within the Bureau, key directives were not filed in official records as the law required (Select Committee to Study Governmental Operations with Respect to Intelligence Activities 1976, p. 355). Thus a power-wielding configuration within the FBI weaponized the national security black box to disable the information access of checkers in Congress, in other parts of the executive branch, and even in the Bureau itself.

Moreover, when checkers sought to scrutinize the FBI's activities, Director Hoover would leverage a secret "archive of blackmail material" to "discredit and destroy his adversaries....With these tactics, Hoover frequently defied Congress and the eight Presidents under whom he served" (Schmautz 1992). The checkers experienced all four chilling-effect factors: observation and personalization through the targeting itself, uncertainty about Hoover's possession and use of these materials, and the power conferred on Hoover by this information asymmetry and the FBI resources at his disposal. These practices chilled checkers into compliance with Hoover, leading to the routine approval of his budget and the expansion of his authority.

During this period, the authoritarian practices of the power-wielders were relatively labor-intensive and technologically constrained. Although COINTELPRO was effective in achieving power-wielders' immediate objectives, the reliance on human agents generated significant operational friction. Furthermore, the power-wielders' attempts to disable information access were limited by the physical stor-

age of the information in paper dossiers that, if uncovered, could be understood by the checkers and the forum. This paper trail, though imperfect, was instrumental for launching legal challenges and assigning accountability.

The War on Terror

During the 21st century, as the over-arching security paradigm in Western formal democracies shifted from the Cold War to the "Global War on Terror" (Buzan 2006; Stohl 2008), authoritarian practices continued to evolve within democratic systems. Among the most consequential developments of this period was power-wielders' rapid adoption of digital technologies. The 2013 Snowden leaks provided unprecedented insights into the practices of national security agencies. The leaked documents revealed programs like the NSA's PRISM and GCHQ's TEMPORA, whereby power-wielders were conducting mass surveillance on the states' own citizens (Hu 2015). The surveillance in itself was illiberal as opposed to authoritarian (Glasius and Michaelsen 2018), but in disabling both checkers' and the fora's access to information about the surveillance programs, the power-wielders engaged in authoritarian practices.

Glasius and Michaelsen (2018) have discussed several instances of power-wielders disabling access to information about these mass surveillance programs; we seek to expand their analysis in two ways. First, whereas their examples involve officials lying to Congress, even executive checkers experienced disabled information access, as can be illustrated by an example from the United Kingdom. Chris Huhme, who was a cabinet minister and member of the UK's National Security Council (NSC) between 2010 and 2012, has stated that at no point during his tenure was the Cabinet or the NSC briefed on the full extent of GCHQ's digital surveillance operations by intelligence officials. As Huhme later noted, "this lack of information, and therefore accountability, is a warning, that the supervision of our intelligence services needs as much updating as their bugging techniques" (Hopkins and Taylor 2013). Without whistleblowers such as Snowden, this information would have remained hidden not only from the forum, but also from the checkers supposed to operate inside the national security black box.

Second, national security power-wielders not only lied to checkers (directly or by omission), they also used chilling effects to manage the boundary of the national security black box. The "offensive" nature of operations against checkers seeking to penetrate the national security black box is illustrated by the 2014 revelation that elements of the CIA covertly surveilled the Senate Intelligence Committee, "secretly removing documents from the Committee [and] searching computers used by the Committee" (Electronic Privacy Information Center 2014; Feinstein 2014). As we will discuss below, these forms of interference, though secret at the time and not designed to create chilling effects, have likely contributed to chilling effects once revealed to the Senate. A black box-reinforcing tactic that was aimed at chilling was the attempt "to intimidate congressional investigators by requesting an FBI inquiry into their conduct" (Electronic Privacy Information Center 2014; Fe-

instein 2014). An FBI inquiry would have made the Committee members' practical subordination to the CIA and FBI salient, as well as leading the members to sense observation. The request itself constituted a personalized threat, and it is possible to list multiple sources of uncertainty for the committee members. Some of the four chilling effect factors were present just because of the request for the inquiry, and the rest would have been added had the inquiry been opened. The CIA officials attempted to chill the Committee members into compliance and conformity with the norms those officials wished to govern the national security black box.

The disclosures of digital mass surveillance had chilling effects on the public (Guo 2016; Penney 2016, 2025). As the chilling effect of awareness of surveillance was not part of the intent of those who created the mass surveillance programs, this is an example of the illiberal practice of mass surveillance eventually affecting democracy in a similar way to an authoritarian practice. In the case of targeted surveillance of checkers, the surveillance itself is already authoritarian. When it is revealed to the checkers, it intensifies any chilling effect that the power-wielders' known past conduct has already created.

AI and Authoritarian Practices

Power-wielders of today's security state are adopting and perhaps creating AI tools to enhance their instruments of control. Some of this activity meets our criteria for authoritarian practices.

The Evolving National Security Black Box

Power-wielders, including private-sector contributors to the security state, continually refashion the national security black box to disable access to information about their AI-enabled activities. Events in the United States exemplify how power-wielders in formal democracies attempt to neutralize application of the legal frameworks designed to guarantee public access to government records. Essential to the US public records framework is the Freedom of Information Act (FOIA). But "as AI increasingly permeates the government's administrative agencies, it exposes the profound shortcomings of the FOIA in its current form, particularly in ensuring transparency around agency decision-making and in effectively informing the public about decisions driven by AI" (Capaldi 2025). The American Civil Liberties Union (ACLU) has an ongoing lawsuit against the National Security Agency (NSA) under FOIA.²⁵ to release key information and discover what, if any, safeguards exist to protect the civil liberties of citizens (Matishak 2023) as the NSA adopts AI. In March 2024, the ACLU requested documents such as studies, roadmaps, and reports detailing how the NSA deploys AI. The government denied the request; in the resulting ongoing litigation, the ACLU is seeking to compel disclosure (American Civil Liberties Union and American Civil Liberties Union Foundation 2024; American Civil Liberties Union 2024a,b).

State-corporate power-wielding configurations use the intellectual property regime to disable access to information about not only the behavior of the corporations, but also the

use of AI in government functions. FOIA's Exemption 4 protects "trade secrets and commercial or financial information obtained from a person [that is] privileged or confidential" (United States Department of Justice, Office of Information Policy 2004). In the present era of AI governance, this exemption is a formidable shield that disables information access. As governments rarely build their own sophisticated AI systems "from scratch," instead opting to purchase or license products from private vendors, the underlying source code, weighting variables, and training datasets are legally classified as corporate property. When citizens, journalists, or civil rights and liberties organizations demand transparency regarding how a predictive policing algorithm assigns risk, or how a welfare distribution algorithm denies benefits, the state-corporate configuration prioritizes the property rights of the corporate vendor over the public's rights to due process and transparent governance of the systems affecting people's lives (Bloch-Wehba 2021; Rashkovich 2019). Although in this paper we emphasize practices originating in the core of the national security apparatus, authoritarian disabling of information access extends far beyond the security agencies and into the rest of the administrative state.

Disabling the Forum's Voice

We have established that state and state-corporate power-wielders are engaging in authoritarian practices of disabling fora's access to information about likely illiberal AI-enabled surveillance practices. But are the AI-enabled surveillance practices themselves authoritarian? By examining public statements by national security officials in formal democracies about how their agencies use or will use AI, we can reasonably infer that national security power-wielders intend to use AI to disable citizens' voice both by actually intervening on the basis of AI predictions and by designing systems of surveillance that chill citizens' behavior into conformity and compliance with norms dictated by the power-wielders.

The Directorate of Digital Innovation, established in 2015, is a specialized branch within the Central Intelligence Agency (CIA) to accelerate the integration of digital capabilities into intelligence operations. According to Andrew Hallman, then the Deputy Director for Digital Innovation, the adoption of new data analysis techniques including those based on deep learning significantly improved the agency's capacity to produce "anticipatory intelligence" that can preempt "social unrest" (Konkel 2016). Similarly, Dawn Meyerriecks, former Deputy Director of the CIA's Directorate of Science & Technology, boasted of the agency's hundreds of AI-related projects intended to "allow us to be more predictive with what's about to go down.... We have that in pockets" (Tucker 2017).⁷

The main UK intelligence services are also integrating AI systems for intelligence collection and action. An article published by the Financial Times, which includes an

⁷It is interesting, though not directly pertinent to the present point, that in the same 2017 interview, Meyerriecks indicated that algorithmic opacity presented a barrier to convincing officials to accept conclusions reached with AI: "you can't go to leadership and make a recommendation based on a process that no one understands."

interview of a senior director of a team focused on developing GCHQ's "next generation artificial intelligence capabilities," provides a rare insight into the systems and practices of the UK's largest intelligence service (Bond 2019). Among the tools and use cases discussed is a system that can process "enormous biometric datasets" to identify behavioral patterns, as well as an advanced facial recognition model that can "identify people from the small 'envelope' of skin beneath the eyes," making many forms of concealment useless. Although we cannot know what the state of the art is today, we should assume that the programs have advanced significantly in the seven years following this disclosure.

A main purpose of identifying behavioral patterns and individuals is to allow officials to intervene, using AI further or not, to stop undesired behavior. But noting the uncertainty about the present state of the art leads us to consider the chilling effects that national security power-wielders can create even without physically intervening on the basis of AI predictions, and perhaps even without using AI as much as we might assume. When evidence from immediately prior surveillance programs together with general knowledge of advances in computing gives the public reason to suspect that surveillance technology may be quite advanced, speech and behavior can be chilled even in the absence of the exact technology imagined. The demonstration of enough technical capabilities to create such a chilling effect can therefore constitute an AI-enabled authoritarian practice.

There is no need to rehearse Penney's extensive discussion of the chilling effects of AI surveillance. The main contribution of this subsection is to show that the stated intentions of some national security officials, when viewed in light of the historical cases from prior sections, clearly include disabling the forum's voice and are therefore authoritarian. Most remarkably, it appears that some national security officials think in the same terms as critical scholars like Penney. In a recent interview with former Prime Minister Tony Blair, UK Home Secretary Shabana Mahmood said "AI and technology can be transformative to the whole of the law and order space," and "When I was in justice, my ultimate vision for that part of the criminal justice system was to achieve, by means of AI and technology, what Jeremy Bentham tried to do with his Panopticon. That is that the eyes of the state can be on you at all times" (Ashworth-Hayes 2026). If the AI Panopticon were used only for a small part of the criminal justice system involving convicted violent criminals, it might raise due process and human rights concerns, but its possible authoritarianism would be only ancillary. Given Mahmood's reference to the "whole of the law and order space," however, and the gradual expansion of this "space" to encompass more and more areas of public life, the AI Panopticon should be understood as a model that some security officials would like to use for governing entire citizenries, using the uncertainty, observation, and authority factors to chill populations into conformity.

Leaked documents from the Snowden era revealed that GCHQ has a Human Science Operations Cell inside its secret Joint Threat Research Intelligence Group (Greenwald 2014) which worked with academic psychologists to develop "strategic influences" that condition responses of

"obedience"—"where an individual submits to, or complies with, an authority figure"—and "conformity"—"whereby an individual's beliefs, feelings, and behaviors, yield to those norms of a social group" (Dhami 2011). These documents provide some of the clearest evidence that national security officials intentionally develop chilling techniques. This evidence of intent is important background to the next subsection on checkers.

Chilling Checkers

The rest of this section is devoted to power-wielders' use of AI to disable checkers' voice and access to information. One can understand the AI-enabled authoritarian disabling of access to information by applying Deeks's descriptions of the algorithmic black box and double black box problems to cases in which a power-wielder uses the double black box for this purpose. The basic ideas are simple:

The use of algorithmic black boxes in national security operations makes it harder for the public and government officials themselves to assess the operations' domestic and international legality...will complicate the government's own ability to pursue competent policies, as well as its ability to understand after the fact how well it performed...[and will] foster confusion about who to hold responsible for an algorithm's mistake. Because the systems are complex and opaque, they provide "cover" for human decision makers, who can point to the systems as the cause of flawed outcomes. (Deeks 2025, pp. 97-101)

Our focus in this subsection will instead be on the disabling of checkers' voice through chilling effects; our discussion of the algorithmic and double black box will concern its contribution to chilling effects. In general, black boxes contribute via the uncertainty factor.

Chilling Through AI Surveillance Modern enterprises and governments increasingly possess the capacity—afforded largely by recent advancements in AI—to ingest, analyze, and interpret vast streams of employee telemetry in real time (Ferreira and Gromova 2026; Macnish 2025). These capabilities extend beyond the rudimentary tracking of keystrokes and active hours to the semantic analysis of internal communication for emotional sentiment and the continuous vetting of personnel to detect security risks (Majumdar 2026). In the UK, AI-augmented digital surveillance, or "bossware," is an increasingly common feature of the workplace (Watch 2024; Bunwaree et al. 2025; Munn 2024). A recent study conducted by the Chartered Management Institute found that a third of UK employers are now using "bossware" technology to track the workers' activity (Booth 2025). Across the private sector, AI-enabled workplace surveillance is a widely contested issue, with multinational corporations such as Amazon deploying deep-learning based tools for "productivity optimization" and asset protection (Amazon UK 2023) while workers criticize these "intrusive algorithms" (Sainato 2024) which create a "pervasive sense of workplace surveillance" (Spilda et al. 2024).

Employers' use of AI-enabled surveillance is not limited to the well-publicized case of low-wage warehouse workers. There is a growing supply of, and demand for, AI-powered semantic employee surveillance systems (Andrejevic 2024), despite controversy (Carnegie 2023; Wilkins 2026). Offering back-end integration into widely-used communication applications such as Slack and Microsoft Teams, companies like Aware and Teramind are providing major employers with AI tools that monitor internal communications and analyze employee sentiment (Field 2024; Watch 2024). When employees know that every message is scanned by instruments of their employer for dissatisfaction and "toxicity" (Field 2024), they experience an admixture of observation, uncertainty, personalization, and authority. Amba Kak, Executive Director of the AI Now Institute, said this state of affairs "results in a chilling effect on what people are saying in the workplace" (Field 2024).

Here, we are interested in the potentially authoritarian character of workplace surveillance targeting public servants who might serve as checkers. In a recent report published by UNISON, one of Europe's largest trade unions, which represents public services workers, trade representatives expressed concern about how "new technological tools including use of AI, are allowing employers to extend their monitoring and surveillance of workers into new areas of their lives" (UNISON 2025). One such tool used widely across the UK Civil Service is Microsoft Purview (Government Cyber Unit 2026), whose "Insider Risk Management" software "integrates machine learning-based detection controls, dynamic protections, and advanced privacy controls to manage and mitigate insider risks" such as data leakage and security violations (Borges 2025). While Microsoft emphasizes the "privacy by design" architecture, which includes the pseudonymization of users, the underlying capability would provide power-wielders inside the Civil Service with near-total, correlated visibility into the digital actions, communications, and relationships of civil servants. Whether or not Microsoft's privacy safeguards function as advertised, the salient fact for chilling effects is what civil servants perceive as the system's reach. A capability with this scope, deployed by an authority over personnel who cannot inspect how it weighs their behavior, instantiates all four of Penney's chilling factors.

In the US, some of the most advanced forms of employee surveillance exist within the national security and intelligence establishments. One purpose of these tools is to identify the microscopic behavioral signals that precede espionage, sabotage, data exfiltration, or unauthorized leaks (Cybersecurity and Infrastructure Security Agency 2022). Intelligence agencies in the US—like their UK counterparts—now rely on complex User and Entity Behavior Analytics to guard against "insider threats." (Cybersecurity and Infrastructure Security Agency 2022). Forcepoint, a major contractor for the US Department of Defense, was awarded an \$89 million contract by the Defense Information Systems Agency for a fully integrated enterprise user activity monitoring solution (Forcepoint 2022). Forcepoint provides "Risk-Adaptive Protection" and "Data Detection & Response," "with AI-driven accuracy," tracking

hundreds of unique "Indicators of Behavior" continuously in real time (Forcepoint 2025). Forcepoint then uses its "AI mesh" to distinguish abnormal behavior, generate alerts, and recommend "automated remediation steps to help security teams respond before data leave your control." (Forcepoint 2026). Another government contractor, General Dynamics Information Technology (GDIT), which boasts of setting the "gold standard across the intelligence community," claims to have transformed insider threat detection with generative AI. Using decades of research on Insider Threat Risk Modeling and psychological profiling, GDIT uses generative AI to "analyze and score authored communications, alerting analysts to employees' concerning behaviors" such as "disgruntlement" (General Dynamics Information Technology 2026).

The deployment of AI-powered "bossware" and advanced insider threat detection systems within government and national security environments can be understood as engineering an architecture of uncertain (because of algorithmic black boxes) and personalized observation on behalf of an authority. This architecture, with all four chilling factors, is likely to produce conformative chilling effects, creating an environment wherein checkers are prevented from speaking or acting outside the power-wielders' norms. Whistle-blowing has historically been an essential means for checkers inside the national security black box to voice concerns regarding the misconduct of power-wielders (Pozen 2013). But when checkers are aware that every digital interaction is being semantically parsed for signs of dissatisfaction or deviance, the risk (either rationally calculated or instinctively felt) associated with gathering information or planning a disclosure may become psychologically and professionally prohibitive. The pervasive threat of algorithmic observation may chill potential whistle-blowers into self-censorship and strict conformity, thereby disabling their voice (and hence other checkers' access to information).

Chilling Through Automation Bias The authority factor in the chilling effects discussed so far has simply been the authority of the officials deploying an algorithm. We must also consider the way compliance is produced through the authority of algorithms themselves. For example, the same checkers who might resist a power-wielder's bare assertion may defer to that very same assertion when it arrives in an AI system's output. Automation bias and the cognitive atrophy resulting from sustained algorithmic delegation should be viewed not always as failure modes of human-AI interaction, but as affordances that power-wielders can cultivate to sabotage accountability. Especially when combined with the opacity of the algorithmic and double black box, algorithmic authority can create chilling effects converting checkers from auditors of the power-wielders' conduct into ratifiers of it.

Automation bias in general is the tendency of human decision-makers to "ascribe to an automated system greater trust in its analytical capabilities than in their own" (Chesterman 2021, p. 68). It was first documented systematically in aviation cockpits (Mosier et al. 1998; Skitka, Mosier, and Burdick 1999) and has since been observed across medical,

judicial, military, and welfare-administration contexts (Godard, Roudsari, and Wyatt 2012; Parasuraman and Manzey 2010). In bureaucratic contexts, the impulse to follow a computer's recommendation functions as a "heuristic replacement" for vigilant information-seeking and processing, potentially turning an algorithmically generated suggestion into a "trusted final decision" (Citron 2008, p. 1272). Crucially for our purposes, this form of bias is not eliminated by expertise and is resistant to simple training interventions (Parasuraman and Manzey 2010). Even in cases where checkers are formally instructed to scrutinize algorithmic outputs, empirical surveys of human-in-the-loop oversight regimes find that humans are systematically unable to perform the functions that the regimes require (Green 2022). Recent experimental work has advanced the understanding of automation bias into the AI era. Using the Cognitive Reflection Test, Shaw and Nave find evidence of what they call "cognitive surrender...where the user relinquishes cognitive control and adopts the AI's judgment as their own" (Shaw and Nave 2026). This finding suggests that the influence operates beneath the level of conscious deliberation that conventional human-in-the-loop regimes presuppose. For power-wielders that have incorporated AI into their operational structure, this cognitive bias is not necessarily a problem to be solved, but may be a resource to be exploited. Indeed, power-wielders have both an incentive and the capacity to make checkers dependent on AI systems in ways that produce conformity.⁸

In some contexts, the deference that automation bias produces is best understood as conformity to an *algorithmic authority* rather than generic trust. This deference manifests a distinct mode of legitimacy that operates alongside and increasingly in place of conventional institutional authority. According to Pasquale, "Authority is increasingly expressed algorithmically" (Pasquale 2015). Lustig, meanwhile, defines algorithmic authority as the power of algorithms "to direct human action and to verify information, in place of relying exclusively on human authority" (Lustig 2022, p. 2). Similarly, Penney argues that algorithms themselves are "a source of authority to which people tend to defer, and thus conform with, as they would to more traditional forms of authority" (Penney 2025, p. 81). These observations are highly significant for our analysis of authoritarian practices, as they give a political charge to the act of deference. A checker who distrusts a power-wielder may nevertheless extend legitimacy to the power-wielder's actions, if the outcome is perceived as being the impersonal product of an algorithmic system. This case adds complexity to Penney's original notion of authority. Power-wielders who have lost or otherwise lack institutional authority can recover or obtain it by routing their decisions through systems that carry algorithmic authority instead. When coupled with the opacity of the algorithmic black box, this substitution becomes self-protecting,

⁸The use of automation bias for chilling shows that power-wielders cannot use all of the uncertainty affordances of AI concurrently with the power/authority affordances: for example, getting a checker to defer to an algorithm because it is an algorithm requires letting the checker know that an algorithm is being used.

as the basis on which algorithmic authority might be revoked is itself unavailable for inspection.

Sustained reliance on AI systems can produce cognitive atrophy in the checkers themselves, eroding the capacities on which any future independent scrutiny would depend. Despite making recent headlines (BBC News 2025; Lufkin 2026; Swarz 2026), this is not an entirely novel issue. In the 1980s, Bainbridge observed that when automated systems handle routine operations, the human operator's manual and analytical skills decay because they are no longer exercised, with the result that the operator is least equipped to intervene at exactly the moments when intervention is most needed (Bainbridge 1983). Further research by Parasuraman and Manzey has found that this complacency effect is observable in both naive and expert test subjects and "cannot be overcome with simple practice" (Parasuraman and Manzey 2010, p. 381). More recent studies have extended our understanding of cognitive atrophy into the domain of today's AI, with Kosmyna and colleagues describing the longitudinal erosion observed in LLM-assisted users as the "accumulation of cognitive debt" (Kosmyna et al. 2025). If "meaningful oversight requires a person with authority and competence to review a decision – including having access to 'all the relevant data'" (Chesterman 2021, p. 69), cognitive atrophy can contribute to accountability sabotage. Moreover, as laws, technical standards, and operational practices governing national security AI grow more intricate, checkers progressively lose both the motivation and the trained capacity to interrogate them, retaining the procedural fact of oversight while its substantive content migrates into systems they cannot read. As Citron warned of an earlier generation of automated administrative systems, the cumulative consequence is that human oversight is "effectively rendered ineffective" (Citron 2008, p. 1277).

Automation bias and cognitive atrophy prevent checkers from acting in deviant ways and condition them to act in cooperation with power-wielders by default. The psychological submission to algorithmic authority has the same effect that, in previous eras, required personalized intimidation like Hoover's blackmail archive or the CIA's interference with the Senate Intelligence Committee. Whereas Hoover had to cultivate fear in each individual checker, the algorithmic substitution of authority and the degradation of independent judgment can generalize this chilling effect across entire checker substrata, without the political risk that any specific act of intimidation would generate. The result can be a population of checkers who are not coerced into conformity, but who, in the relevant sense, cannot help complying.

Conclusion

This paper has examined one kind of AI-related power dynamics, namely AI-enabled state or state-corporate authoritarianism understood as AI-enabled practices of accountability sabotage. We have demonstrated several points that should interest readers approaching "AI ethics and society" from a variety of backgrounds and disciplinary perspectives. First, we have established that the history of national security practices in formal liberal democracies should lead one to expect that AI adoption in these regimes will not merely

create new gaps in accountability, but will be used by power-wielders for authoritarian practices. This point is relevant for political scientists who had previously been conceiving AI-related authoritarianism in terms of formally undemocratic regimes, critical scholars who had studied AI-related power inequalities in formal democracies but had not yet identified the specifically authoritarian political phenomenon, and computer scientists and engineers who had considered accountability issues in terms of gaps rather than sabotage.

Second, we have updated Glasius and Michaelson's 2018 treatment of digital authoritarian practices to cover novelties of contemporary AI. Today's AI models' accuracy, actual and perceived, and their opacity both contribute to the software's role in authoritarian practices. "AI ethics and society" scholars specializing in issues related to the social construction of "accuracy," false perceptions of accuracy, and the social consequences of technical opacity can now understand these elements as authoritarian affordances.

In applying the frameworks of the double black box and chilling effects as conformity effects, we have gone beyond the work of both Deeks and Penney by combining the frameworks with each other, with Glasius's conception of authoritarianism, and with supplementary evidence. Thus we showed that the double black box of national security contributes to the uncertainty and power factors in chilling effects, with an additive property in the case of the multiple sources of uncertainty from the two black boxes. And AI-enabled chilling effects are seen to afflict checkers of the national security apparatus. Our discussion of authoritarian practices directed at checkers raises difficult questions for procedurally-oriented "AI accountability" commentators whose proposals of clever governance mechanisms fail to account for the ruthlessness and ingenuity of power-wielders vis-à-vis checkers.

We have introduced useful frameworks and indicated some previously neglected considerations without attempting a comprehensive survey of any of the topics we have discussed. Directions in which the work in this paper could be extended include—in addition to thorough studies of more cases—study of AI-enabled corporate authoritarianism less directly mediated by the state, so that it is not "state-corporate authoritarianism," and study of AI-enabled authoritarianism in transnational governance arrangements involving various configurations of state and corporate actors. In the transnational arena, there are few formal mechanisms designed to make power-wielders based in more powerful countries accountable to people they affect in less powerful countries, and therefore the procedural rule-breaking aspect of accountability sabotage is less salient, though Glasius shows that transnational authoritarianism is yet a relevant concept. "AI ethics and society" scholars interested in issues of global inequality but not in distinguishing transnational authoritarianism from other power dynamics may nevertheless find value in our contributions regarding formal liberal democracies, since the status of authoritarianism in these regimes is plausibly related to the states' conduct in foreign affairs.

We close by noting limitations in our approach that derive from our choice of conception of authoritarianism. In a re-

view of Glasius's 2023 book, Salvador Santino F. Regilme Jr. raises a problem of researcher positionality and moreover writes

the focus on practices as units of analysis has the potential to overlook structural and systemic factors that generate and sustain such practices (Regilme 2024).

Also regarding systemic factors, we can observe that in searching for conceptual frameworks to organize our writing about AI-related social issues, we are biased in favor of those that we can easily understand and quickly and confidently apply. Which ones those are is of course determined by numerous social factors.

Finally, the role of AI in social control may increase the overlap between illiberal and authoritarian practices or even render the conceptual distinction irrelevant. "AI ethics and society" scholars now understand that part of the currently dominant AI agenda is to make people's behavior more predictable rather than just predict it. As Penney puts the point in terms of his theory,

Chilling effects [including those created with AI] thus produce individuals and populations that are more conforming, compliant, and easier to manage, organize, and sort into hierarchies – and, most importantly for the purposes of our discussion, easier for AI to analyze, predict, and subsume within the authoritarian dominion of AI capitalism, and the surveillance systems and other mass tracking and enforcement technologies that will follow (Penney 2025, p. 178).

As we noted, Penney does not discuss the concept of authoritarianism at length, and in particular does not distinguish it from illiberalism. At some point in the program of social engineering he is describing, the question of whether the main effect or purpose of a power-wielder's action is to undermine personal autonomy or to sabotage accountability will have lost its meaning.

Acknowledgments

The authors thank Jeba Sania for information about the literature on digital authoritarianism, Huw Roberts for comments on a draft of this paper, and Luke Kemp for insightful discussions throughout the development of the paper. This work was supported by the Oxford Martin AI Governance Initiative, and by the Future of Life Institute (FLI) through its grant program on mitigating AI-driven power concentration. The FLI grant is administered by the Berkeley Existential Risk Initiative. FB was supported by a grant from the Cosmos Institute.

References

- Amazon UK. 2023. Written evidence from Amazon UK (ULM0119). UK Parliament Business, Energy and Industrial Strategy Committee. <https://committees.parliament.uk/writtenevidence/120164/default/>. Submitted to the Post-pandemic economic growth: UK labour markets inquiry. Accessed: 2026-05-08.
- American Civil Liberties Union. 2024a. ACLU v. NSA – FOIA Lawsuit Seeking Records About the NSA's Use

- of Artificial Intelligence. <https://www.aclu.org/cases/aclu-et-al-v-nsa-et-al-foia-lawsuit-seeking-records-about-the-nssas-use-of-artificial-intelligence>. Status: Ongoing. Last updated May 13, 2024.
- American Civil Liberties Union. 2024b. Request Under Freedom of Information Act Concerning the National Security Agency's Use of Artificial Intelligence. FOIA Request to the National Security Agency, Department of Defense, and Office of the Director of National Intelligence. Litigation filed April 25, 2024, in the U.S. District Court for the Southern District of New York.
- American Civil Liberties Union; and American Civil Liberties Union Foundation. 2024. Complaint for Injunctive Relief: American Civil Liberties Union v. National Security Agency, et al. United States District Court for the Southern District of New York, Case No. 1:24-cv-03153.
- Andrejevic, M. 2024. Automated monitoring in the workplace: the devolution of recognition. *International Journal of Communication*, 18: 3205–3211.
- Ashworth-Hayes, S. 2026. Criminals to be Stopped by Minority Report-Style Policing Plans. The Telegraph. <https://www.telegraph.co.uk/news/2026/01/18/criminals-stopped-minority-report-style-policing-plans-ai/>.
- Bainbridge, L. 1983. Ironies of Automation. *Automatica*, 19(6): 775–779.
- BBC News. 2025. Are these AI prompts damaging your thinking skills? <https://www.bbc.co.uk/news/articles/cd6xz12j6pzo>. Accessed: 2026-05-13.
- Bloch-Wehba, H. 2021. Transparency's AI Problem. Knight First Amendment Institute at Columbia University. <https://knightcolumbia.org/content/transparencys-ai-problem>. Accessed: 2026-05-15.
- Bond, D. 2019. Inside GCHQ: The Art of Spying in the Digital Age. Financial Times. <https://www.ft.com/content/cc68ffc-7c1e-11e9-81d2-f785092ab560>. Accessed: 2026-08-07.
- Booth, R. 2025. A third of UK firms using 'bossware' to monitor workers' activity. The Guardian. <https://www.theguardian.com/world/2025/sep/14/uk-firms-bossware-monitor-workers-activity>. Accessed: 2026-04-26.
- Borges, N. 2025. Mitigating insider risks in the age of AI with Microsoft Purview Insider Risk Management. Microsoft Security Community Blog. <https://techcommunity.microsoft.com/blog/microsoft-security-blog/mitigating-insider-risks-in-the-age-of-ai-with-microsoft-purview-insider-risk-ma/4396057>. Accessed: 2026-04-26.
- Bovens, M. 2007. Analysing and Assessing Accountability: A Conceptual Framework. *European Law Journal*, 13(4): 447–468.
- Bunwaree, T. S.; Stawarz, K.; Collins, P.; and Gould, S. J. 2025. Boss is aWare—Are you? Employee comprehension and legal awareness of workplace monitoring. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–19.
- Buzan, B. 2006. Will the 'global war on terrorism' be the new Cold War? *International affairs*, 82(6): 1101–1118.
- Capaldi, R. L. 2025. FOIA and the Use of AI in Government: Freedom of Information or an Empty Promise of Openness? *Journal of the National Association of Administrative Law Judiciary*, 45(2): 25–53.
- Carnegie, M. 2023. The Creepy Rise of Bossware. Wired. <https://www.wired.com/story/creepy-rise-bossware/>. Accessed: 2026-08-07.
- Carpenter, D. 2026. Authoritarian State Conversion: America's Past and the Trumpian Future. *Political Science Quarterly*, 141(1): 39–49.
- Chesterman, S. 2021. *We, the Robots?: Regulating Artificial Intelligence and the Limits of the Law*. Cambridge University Press.
- Citron, D. K. 2008. Technological Due Process. *Washington University Law Review*, 85(6): 1249.
- Crawford, K. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.
- Cybersecurity and Infrastructure Security Agency. 2022. Insider Threat Mitigation Guide. https://www.cisa.gov/sites/default/files/2022-11/Insider%20Threat%20Mitigation%20Guide_Final_508.pdf. Accessed: 2026-05-20.
- Deeks, A. S. 2025. *The Double Black Box: National Security, Artificial Intelligence, and the Struggle for Democratic Accountability*. New York, NY: Oxford University Press.
- Dhami, M. K. 2011. Behavioural Science Support for JTRIG's Effects and Online HUMINT Operations. The Intercept. <https://theintercept.com/document/2015/06/22/behavioural-science-support-jtrig/>. Document leaked and published online on June 22, 2015. Accessed: 2026-04-29.
- Dragu, T.; and Lupu, Y. 2021. Digital Authoritarianism and the Future of Human Rights. *International Organization*, 75(4): 991–1017.
- Electronic Privacy Information Center. 2014. EPIC v. CIA – CIA Spying on Congress. <https://epic.org/documents/epic-v-cia-cia-spying-on-congress/>. Accessed: 2026-04-17.
- Erixon, F.; and Lee-Makiyama, H. 2011. Digital Authoritarianism: Human Rights, Geopolitics and Commerce. European Centre for International Political Economy. <https://ecipe.org/publications/digital-authoritarianism-human-rights-geopolitics-and-commerce/>.
- Feinstein, D. 2014. Statement on CIA Search of Intelligence Committee Computers. Congressional Record, 113th Congress, 2nd Session, Vol. 160, No. 39. Speech delivered on the floor of the United States Senate detailing the CIA's search and the subsequent crimes report filed against committee staff.
- Feldstein, S. 2021. *The Rise of Digital Repression: How Technology is Reshaping Power, Politics, and Resistance*. New York, NY: Oxford University Press.
- Ferreira, D. B.; and Gromova, E. A. 2026. The Bossware Era and the e-Panopticon: Current Technologies and Legal Challenges. *Industrial Law Journal*, 55(1): 181–212.

- Field, H. 2024. How Walmart, Delta, Chevron and Starbucks are using AI to monitor employee messages. CNBC. <https://www.cnbc.com/2024/02/09/ai-might-be-reading-your-slack-teams-messages-using-tech-from-aware.html>. Accessed: 2026-04-24.
- Forcepoint. 2022. Forcepoint Federal Awarded \$89M Department of Defense Contract For User Activity Monitoring Enterprise Solution. <https://www.forcepoint.com/newsroom/2022/forcepoint-federal-awarded-89m-department-defense-contract-user-activity-monitoring>. Accessed: 2026-04-26.
- Forcepoint. 2025. Government Data Security Solutions with Forcepoint. <https://www.forcepoint.com/sites/default/files/2025-06/brochure-government-data-security-solutions-with-fp-en.pdf>. Accessed: 2026-04-26.
- Forcepoint. 2026. Data Detection and Response (DDR). <https://www.forcepoint.com/product/data-detection-response-ddr>. Accessed: 2026-04-26.
- Frantz, E.; Kendall-Taylor, A.; and Wright, J. 2020. Digital Repression in Autocracies. V-Dem Users Working Paper No. 27, The Varieties of Democracy Institute (V-Dem), Gothenburg.
- General Dynamics Information Technology. 2026. Transforming Insider Threat Detection with Generative AI. <https://www.gdit.com/perspectives/latest/transforming-insider-threat-detection-with-generative-ai/>. Accessed: 2026-04-26.
- Glasius, M. 2018. What authoritarianism is... and is not: a practice perspective. *International affairs*, 94(3): 515–533.
- Glasius, M. 2023. *Authoritarian Practices in a Global Age*. Oxford, England: Oxford University Press.
- Glasius, M.; and Michaelsen, M. 2018. Authoritarian Practices in the Digital Age— Illiberal and Authoritarian Practices in the Digital Sphere — Prologue. *International Journal of Communication*, 12: 3795–3813. Special Section: Authoritarian Practices in the Digital Age.
- Goddard, K.; Roudsari, A.; and Wyatt, J. C. 2012. Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators. *Journal of the American Medical Informatics Association*, 19(1): 121–127.
- Government Cyber Unit. 2026. Use Cyber Uplift's Microsoft Security Optimisation Capability. <https://www.security.gov.uk/services-resources/cyber-services-government/cyber-uplift/microsoft-security-optimisation/>. UK Government Security. Accessed: 2026-09-10.
- Grant, R. W.; and Keohane, R. O. 2005. Accountability and Abuses of Power in World Politics. *American Political Science Review*, 99(1): 29–43.
- Green, B. 2022. The Flaws of Policies Requiring Human Oversight of Government Algorithms. *Computer Law & Security Review*, 45: 105681.
- Greenwald, G. 2014. How Covert Agents Infiltrate the Internet to Manipulate, Deceive, and Destroy Reputations. *The Intercept*. <https://theintercept.com/2014/02/24/jtrig-manipulation/>. Accessed: 2026-04-29.
- Guo, J. 2016. New study: Snowden's disclosures about NSA spying had a scary effect on free speech. *The Washington Post*. <https://www.washingtonpost.com/news/wnk/wp/2016/04/27/new-study-snowdens-disclosures-about-nsa-spying-had-a-scary-effect-on-free-speech/>. Accessed: 2026-04-15.
- Haugeland, J. 1985. *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.
- Hintz, A.; and Milan, S. 2018. Authoritarian Practices in the Digital Age| “Through a Glass, Darkly”: Everyday Acts of Authoritarianism in the Liberal West. *International Journal of Communication*, 12: 21–21.
- Hopkins, N.; and Taylor, M. 2013. Cabinet was told nothing about GCHQ spying programmes, says Chris Huhne. *The Guardian*. <https://www.theguardian.com/uk-news/2013/oct/06/cabinet-gchq-surveillance-spying-huhne>. Accessed: 2026-04-15.
- Hu, M. 2015. Taxonomy of the snowden disclosures. *Wash. & Lee L. Rev.*, 72: 1679.
- Konkel, F. 2016. CIA Can Anticipate Social Unrest ‘Three to Five Days’ Out in Some Cases. Nextgov. <https://www.nextgov.com/digital-government/2016/10/cia-can-predict-social-unrest-three-five-days-out/132102/>. Accessed: 2026-08-07.
- Kosmyna, N.; Hauptmann, E.; Yuan, Y. T.; Situ, J.; Liao, X.-H.; Beresnitzky, A. V.; Braunstein, I.; and Maes, P. 2025. Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task. arXiv:2506.08872.
- Lamensch, M. 2021. Authoritarianism Has Been Reinvented for the Digital Age. Centre for International Governance Innovation. <https://www.cigionline.org/articles/authoritarianism-has-been-reinvented-for-the-digital-age/>. Accessed: 2026-08-07.
- Lufkin, B. 2026. ‘Think outside the bots’: How to stop AI from turning your brain to mush. BBC Future. <https://www.bbc.co.uk/future/article/20260505-how-to-use-ai-without-turning-your-brain-to-mush>. Accessed: 2026-05-13.
- Lustig, C. 2022. Algorithmic Authority: The Case of Bitcoin. arXiv:2201.05939.
- Macnish, K. 2025. Surveillance. *A Companion to Digital Ethics*, 267–279.
- Maerz, S. F. 2026. How practices of digital authoritarianism harm democracy. *Democratization*, 33(1): 163–190.
- Majumdar, P. 2026. The Economics of Distrust: Systemic Risks of Digital Surveillance and Corporate Insecurity in the Modern Enterprise. *Journal of Human Resource Management*, 14(1): 12–21.
- Matishak, M. 2023. NSA, Cyber Command Recently Wrapped Studies on AI Use, Director Says. *The Record from Recorded Future News*. <https://perma.cc/F4ZT-PNTB>. Accessed: 2026-04-22.
- Mosier, K. L.; Skitka, L. J.; Heers, S.; and Burdick, M. 1998. Automation Bias: Decision-Making and Performance in High-Tech Cockpits. *International Journal of Aviation Psychology*, 8(1): 47–63.

- Munn, L. 2024. Expansive and invasive: Mapping the “bossware” used to monitor workers. *Surveillance & Society*, 22(2): 104–119.
- Parasuraman, R.; and Manzey, D. H. 2010. Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors*, 52(3): 381–410.
- Pasquale, F. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.
- Penney, J. 2016. How surveillance harms. Policy Options. <https://policyoptions.irpp.org/2016/12/how-surveillance-harms/>. Accessed: 2026-04-15.
- Penney, J. W. 2025. *Chilling Effects: Repression, Conformity, and Power in the Digital Age*. Cambridge: Cambridge University Press.
- Polyakova, A.; and Meserole, C. 2019. Exporting Digital Authoritarianism: The Russian and Chinese Models. Policy Brief, Democracy and Disorder Series, Foreign Policy at Brookings. https://www.brookings.edu/wp-content/uploads/2019/08/FP_20190827_digital-authoritarianism_polyakova_meserole.pdf. Accessed: 2026-08-07.
- Pozen, D. E. 2013. The Leaky Leviathan: Why the Government Condemns and Condone Unlawful Disclosures of Information. *Harvard Law Review*, 127(2): 512–635.
- Rashkovich, B. 2019. Government Accountability in the Age of Automation. Yale Law School, Media Freedom and Information Access Clinic. <https://law.yale.edu/mfia/case-disclosed/government-accountability-age-automation>. Accessed: 2026-05-15.
- Regilme, S. S. F. 2024. Marlies Glasius: Authoritarian Practices in the Global Age. *New Global Studies*, 18(3): 293–296.
- Sainato, M. 2024. ‘You feel like you’re in prison’: workers claim Amazon’s surveillance violates labor law. The Guardian. <https://www.theguardian.com/us-news/article/2024/may/21/amazon-surveillance-lawsuit-union>. Accessed: 2026-05-08.
- Schmautz, K. A. 1992. J. Edgar Hoover: The Man and the Secrets. *Michigan Law Review*, 90(6): 1812.
- Select Committee to Study Governmental Operations with Respect to Intelligence Activities. 1976. Final Report: Book III: Supplementary Detailed Staff Reports on Intelligence Activities and the Rights of Americans. Technical Report S. Rep. No. 94-755, United States Senate, Washington, D.C. 94th Congress, 2nd Session.
- Shaw, S. D.; and Nave, G. 2026. Thinking—Fast, Slow, and Artificial: How AI is Reshaping Human Reasoning and the Rise of Cognitive Surrender. The Wharton School Research Paper. Available at SSRN: <https://dx.doi.org/10.2139/ssrn.6097646>.
- Skitka, L. J.; Mosier, K. L.; and Burdick, M. 1999. Does Automation Bias Decision-Making? *International Journal of Human-Computer Studies*, 51(5): 991–1006.
- Spilda, F. U.; Brittain, L.; Cant, C.; Cole, M.; Alyanak, O.; Mozzachiodi, R.; and Graham, M. 2024. Fairwork Amazon Report 2024: Transformation of the Warehouse Sector through AI. Fairwork and the Global Partnership on Artificial Intelligence. https://fair.work/wp-content/uploads/sites/17/2024/06/Fairwork-Amazon-Report_June24-2.pdf. Accessed: 2026-05-08.
- Stohl, M. 2008. The global war on terror and state terrorism. *Perspectives on Terrorism*, 2(9): 4–10.
- Swarz, T. 2026. Brain scientist warns we’re heading for AI-fueled dementia crisis. New York Post. <https://nypost.com/2026/04/10/health/brain-scientist-warns-were-heading-for-ai-fueled-dementia-crisis/>. Accessed: 2026-05-13.
- Tucker, P. 2017. What the CIA’s Tech Director Wants from AI. Defense One. <https://www.defenseone.com/technology/2017/09/cia-technology-director-artificial-intelligence/140801/>. Accessed: 2026-08-07.
- UNISON. 2025. Artificial intelligence (AI) and other new technology in the workplace. <https://unison.org.uk/get-help/artificial-intelligence-ai-and-your-rights>. Accessed: 2026-04-26.
- United States Department of Justice, Office of Information Policy. 2004. FOIA Guide, 2004 Edition: Exemption 4. Accessed: 2026-04-22.
- Watch, B. B. 2024. Bossware: The Dangers of High-Tech Worker Surveillance and How to Stop Them. <https://bigbrotherwatch.org.uk/wp-content/uploads/2024/09/BosswareWebVersion.pdf>. Accessed: 2026-08-07.
- Wilkins, J. 2026. AI Is Giving Your Boss Tools to Be More Monstrous Than Ever Before. Futurism. <https://futurism.com/artificial-intelligence/ai-boss-surveillance>. Accessed: 2026-08-07.
- Zuboff, S. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs.