
The Journal of Information Systems Security is a publication of the Information Institute. The JISSec's mission is to significantly expand the domain of information system security research to a wide and eclectic audience of academics, consultants and executives who are involved in the management of security and generally maintaining the integrity of the business operations.

Editor-in-Chief
Gurpreet Dhillon
University of North Texas, USA

Managing Editor
Filipe de Sá-Soares
University of Minho, Portugal

Publishing Manager
Mark Crathorne
ISEG, Universidade de Lisboa, Portugal

Print ISSN: 1551-0123
Online ISSN: 1551-0808
Volume 21, Issue 3

www.jissec.org

WHAT WOULD FOUCAULT SAY ABOUT AI, SURVEILLANCE, AND CYBERSECURITY?

Gurpreet Dhillon

University of North Texas, USA

Michel Foucault's theories of power – particularly the panopticon, governmentality, biopolitics, and resistance – provide a robust lens for analyzing the convergence of artificial intelligence (AI), surveillance, and cybersecurity in contemporary society. These technologies are not merely technical innovations but sophisticated apparatuses of control that extend and intensify modern disciplinary and biopolitical regimes. By automating observation, producing norms, shaping digital subjects, and provoking resistance, AI-driven cybersecurity redefines power in a digital episteme. This editorial essay elaborates on how Foucault might interpret these systems, exploring their mechanisms, implications, and the possibilities for counter-conduct.

I. AI as a Post-Human Panopticon

Foucault's panopticon, introduced in *Discipline and Punish*, describes a mechanism of invisible observation that induces self-discipline. Unlike Bentham's architectural prison, where a central watchtower enabled constant potential surveillance, AI-powered cybersecurity systems create a post-human panopticon that operates without human overseers. Technologies like anomaly detection algorithms, fraud scoring models, and predictive firewalls monitor digital interactions in real time, flagging deviations from "normal" behavior. For instance, machine learning systems in banking detect "unusual" transactions, while network security tools identify "suspicious" IP addresses—all without direct human intervention.

This automation renders power opaque and unaccountable. Users are unaware of the specific criteria or datasets that trigger flags, yet they modify their behavior – avoiding certain websites, using "secure" passwords, or limiting encrypted communications – to evade scrutiny. The invisibility of AI's gaze intensifies the panoptic effect: individuals internalize discipline, assuming they are always being

watched by an impersonal, algorithmic authority. Foucault might argue that this shift from human to machine surveillance marks a new phase of disciplinary power, where code itself becomes the warden, embedding control into the infrastructure of digital life. Moreover, the global scale of AI surveillance – spanning cloud platforms, social media, and IoT devices – extends the panopticon beyond physical spaces, creating a ubiquitous, inescapable network of observation.

2. Algorithmic Governmentality and Biopolitical Control

In his lectures on *Security, Territory, Population*, Foucault described governmentality as the art of managing populations through statistical knowledge, norms, and institutions. AI-driven cybersecurity embodies this concept by transforming security from reactive punishment to proactive risk management. Algorithms aggregate vast datasets – user behaviors, network traffic, device metadata – to calculate risk profiles and assign trust scores. For example, cybersecurity platforms like Microsoft Defender or CrowdStrike use AI to predict threats, while social credit systems, such as those in China, score individuals based on digital footprints, regulating access to services like travel or finance.

This shift aligns with Foucault’s notion of biopolitics, which focuses on managing life at the population level. Rather than punishing specific acts, AI preempts threats by setting thresholds of acceptable behavior. For instance, a user flagged for “excessive” VPN usage might be denied access to a platform – not because of a proven violation but because their behavior deviates from a statistical norm. This preemptive logic governs entire populations, categorizing individuals as “low-risk” or “high-risk” based on opaque metrics. Foucault would likely see this as a biopolitical strategy that regulates life itself, shaping possible futures by controlling access, mobility, and opportunity in digital spaces. The result is a seamless integration of security into everyday life, where compliance becomes a condition of participation.

3. Cybersecurity as a Regime of Truth

For Foucault, power is inseparable from the production of truth. Cybersecurity operates as a regime of truth by defining what constitutes “safe,” “dangerous,” or “vulnerable” in digital interactions. Security vendors, state agencies, and tech platforms wield immense authority to classify behaviors, technologies, and even individuals. For example, antivirus software labels certain files as “malicious,” while platforms like X or Google flag accounts as “suspicious” based on proprietary algorithms. These classifications are not neutral; they reflect the priorities of powerful actors, such as governments criminalizing Tor usage or corporations penalizing peer-to-peer file sharing (e.g., torrenting).

This truth-production shapes societal norms and behaviors. Activities like whistleblowing – exposing state or corporate misconduct – can be framed as “cyberthreats,” as seen in cases like Edward Snowden’s leaks, which were condemned as security violations. Similarly, the use of encryption tools is often stigmatized as “suspicious,” despite being a legitimate privacy practice. Foucault would argue that cybersecurity’s ability to define threats creates a discursive field that legitimizes control. By controlling the narrative around security, these systems normalize certain behaviors (e.g., transparent data sharing) while marginalizing others (e.g., anonymity), reinforcing power through technical expertise. Knowledge, as Foucault noted, is power – and in cybersecurity, the power to classify is the power to control.

4. Docile Digital Selves

Foucauldian disciplinary power produces docile bodies – subjects who conform to norms through subtle, pervasive regulations. Cybersecurity extends this concept to create **docile digital selves**, trained through routines like password policies, phishing awareness training, and two-factor authentication. These practices, often mandated by employers or platforms, instill habits of self-surveillance and risk aversion. For example, employees undergoing phishing simulations learn to scrutinize every email, while users prompted to update passwords regularly internalize cyber hygiene as a personal responsibility.

This discipline operates not through coercion but through voluntary compliance. Users adopt “secure” behaviors – avoiding public Wi-Fi, enabling multi-factor authentication, or limiting data sharing – not out of fear of punishment but to maintain access to digital services. Foucault would see this as a hallmark of disciplinary power: the subject is not forced but shaped into compliance through the design of digital systems. For instance, websites that block access for non-compliant browsers or require CAPTCHA tests enforce micro-regulations that users accept as normal. This disciplinary matrix extends into personal life, where individuals monitor their own digital footprints, wary of being flagged or hacked – thus embedding control into the rhythms of everyday digital existence.

5. Resistance and Counter-Conduct

Foucault’s assertion that “where there is power, there is resistance” finds vivid expression in responses to AI-driven cybersecurity. Counter-conduct – practices that challenge dominant power structures – manifests in technologies and tactics that subvert surveillance. Encryption tools like Signal, VPNs, and decentralized platforms like Mastodon or blockchain-based systems (Web3) enable users to reclaim privacy and reject centralized control. These technologies are not merely practical; they are political acts of defiance against the truth-regimes of cybersecurity, which equate anonymity with threat.

Hactivism, whistleblowing, and data obfuscation further exemplify resistance. Groups like Anonymous use cyberattacks to expose corporate or state misconduct, while whistleblowers like Chelsea Manning reveal the excesses of surveillance systems. Data obfuscation tactics – such as using tracker blockers or generating false digital footprint – disrupt AI’s ability to profile users accurately. Foucault would view these as attempts to re-politicize security, exposing its arbitrariness and challenging its authority. However, resistance is not without risks: counter-conduct often attracts heightened scrutiny, as seen in state crackdowns on VPN usage or the prosecution of hackers. Yet, for Foucault, such risks underscore the vitality of resistance in contesting power’s totalizing grip.

6. Redefining Subjectivity: The Pre-Criminal Identity

AI-mediated cybersecurity redefines subjectivity by judging individuals not by their actions but by their potential risks. Predictive algorithms create **pre-criminal identities**, where users are regulated based on what they might do. For example, a traveler flagged by an AI-based border security system may be detained due to patterns in their travel history, not evidence of wrongdoing. Similarly, social media platforms may suspend accounts for “suspicious activity” detected by algorithms, without transparent justification.

This shift aligns with Foucault’s concept of biopolitics, where power manages populations through probabilistic futures. By constructing subjects as inherently risky, AI cybersecurity systems erode traditional notions of agency and accountability. Individuals are no longer judged by intent or action but by statistical correlations, creating a new form of subjectivity that is both preemptive and dehumanized. Foucault might argue that this represents the ultimate extension of disciplinary power: a regime where the subject is defined not by their present but by their algorithmically predicted future.

A New Digital Episteme

Foucault would likely diagnose AI-powered cybersecurity as a cornerstone of a new **digital episteme** – a regime of knowledge and power defined by algorithmic visibility, internalized digital discipline, and truth-production via technical expertise. These systems automate surveillance, preempt risks, and shape docile digital selves, creating a self-reinforcing matrix of control that operates at a scale and speed beyond human capacity. Yet, as Foucault emphasized, power is never absolute. Resistance through encryption, counter-conduct, and critical discourse offers pathways to challenge this matrix, exposing its contingencies and reclaiming agency.

The stakes of this struggle are profound. AI and cybersecurity do not merely protect digital spaces; they redefine the boundaries of freedom, identity, and possibility. By illuminating the mechanisms of control – through the lens of Foucault’s theories – we can better understand the challenges of navigating a world where power is coded into the algorithms that govern our lives. Resistance, in all its forms, remains essential to imagining alternative digital futures.

Gurpreet Dhillon holds the G. Brint Ryan Endowed Chair of Artificial Intelligence and Cybersecurity at the University of North Texas, USA. He also holds honorary appointments at the University of KwaZulu-Natal, South Africa and Universidade de Lisboa (University of Lisbon), Portugal. Gurpreet earned a PhD from the London School of Economics. He received an Honorary Doctorate from Örebro University, Sweden in 2019. Several of his research papers have been published in FT50 journals. Additionally, he has been featured in the Wall Street Journal, the New York Times, USA Today, Business Week, CNN, NBC News, and NPR.