

UNPACKING AI'S KNOWLEDGE-POWER DYNAMIC: INSIGHTS FROM MICHEL FOUCAULT

Gideon Kato

Department of Philosophy, Augustinian Institute,

Makurdi

Kato.de.elder@gmail.com

Abstract

Artificial intelligence is now one of the biggest agents in contemporary knowledge production and social order construction. AI systems increasingly mediate institutional practice, ranging from diagnostics in healthcare to forecasting in finance, from law enforcement to moderation of truth. Such technologies are far from being neutral but operate to shape reality while inscribing social hierarchies onto digital infrastructures. This is similar to Michel Foucault's thought on power and knowledge. Foucault chronicles prisons, schools, and hospitals as "mechanisms of discipline" that construct knowledge about individuals as thoroughly as they regulate them. This can be applied to AI systems that sort, analyze, and rank populations of people by algorithms. The study unveils three processes of algorithmic power: institutionalizing AI as epistemic authority, data-driven creation of social norms, and automatic marginalization of subjugated knowledges. But in drawing on Foucault's rule that 'where there is power, there is resistance,' we map new spaces for counter-conduct ranging from

algorithmic activism to co-design that subvert AI's disciplinary logic. In conclusion, we make an argument for reconceptualizing AI as an instrument for knowledge-emancipation, setting forth practical principles for democratic governance of algorithms. Through an application of Foucault's critical optic to contemporary technologies, this analysis presents at the same time an alarm about AI's potential to normalize as much as an instruction for cultivating resistant subjectivities in an era of algorithms.

Keywords: Power, knowledge, governance, algorithms, AI

Introduction

Artificial intelligence is now one of the biggest agents in contemporary knowledge production and decision-making. AI systems increasingly mediate institutional practice and knowledge, from diagnostics in healthcare to forecasting in finance, from law enforcement to material moderation. As *Algorithms of Oppression* illustrates, such tools are not neutral but operate to shape reality by inscribing existing social hierarchies onto digital infrastructures (Noble 15). Such an epistemological shift in knowledge authorization and deployment as an apparatus for exerting power is characterized by the development of machine learning systems aiming to “predict” human behavior or ‘discover’ facts about the world from data analysis (Eubanks 47). Such an epistemological shift must be critiqued by analytical

tools that bring political dimensions to epistemic systems.

The concept by Michel Foucault concerning the interplay between knowledge (*savoir*) and power (*pouvoir*) is central to an explanation of AI's role in society. In *Discipline and Punish*, Foucault reveals how institutions like prisons, schools, and hospitals function as "disciplinary mechanisms" that produce knowledge about individuals while simultaneously exercising control over them (Foucault 191). This analysis extends powerfully to AI systems that classify, assess, and sort human populations through algorithmic processes. As Introna maintains, "algorithms are new disciplinary machines" that embody Foucault's panoptic principles in that data is perpetually monitored and comportment is normalized (Introna 118). Further, Foucault's argument that "governmentality" is contemporary power exerted in terms of freedom, not repression, is best enacted in AI systems that observe user comportment when choice is an illusion (Foucault, *Security, Territory, Population* 107).

This essay contends that AI systems constitute a new truth regime that reasserts and reorders relations of power by three principal mechanisms: institutionalized control over knowledge production, algorithmic sorting of social reality, and technical authorization over some epistemologies in favor of others. According to Beer, algorithms acquire their

power not from their computations but from their ability to “define the boundaries of what is knowable” (Beer 9). Following from Foucault’s approach in *The Archaeology of Knowledge*, we need to examine AI’s discursive formations - rules for determining which statement is true in algorithms (Foucault 49). From algorithms for credit scores that normalize uneven economics to facial recognition algorithms that write race-based hierarchies, AI is an apparatus of truth in Foucault’s sense (apparatus or *dispositif*) for late capitalism (Foucault, *Power/Knowledge* 131). By extension of Foucault’s approach, this analysis reveals that AI reproduces historical mechanisms for epistemic domination but involves new mechanisms for digital rule that necessitate new means for resistance and accountability.

Foucault’s Model: Power as Knowledge

Central to Michel Foucault’s theoretical project is the working hypothesis that “power is exercised through the production of knowledge” (Foucault, *Power/Knowledge* 52). This foundational reinterpretation in terms of the power of knowledge overthrows conventional concepts that understand knowledge as autonomous, independent truth. In their place, Foucault demonstrates how regimes of truth have been contingent ones that have been applied to authorize kinds of governance. In *Discipline and Punish*, he reveals how prisons, schools, and hospitals constructed specialist knowledge regimes that in one

blow both typified and regulated populations (Foucault 184). These “disciplinary institutions,” in Foucault’s terminology, functioned by meticulous observing, classifying, and normalizing processes that have suggestive analogs in contemporary AI systems.

The transition to algorithmic knowledge production is central to Foucault’s schema. Schools and hospitals used to be central sites for producing knowledgeable discourse about individuals, but AI systems perform these same acts through computational algorithms. As Zuboff contends in *The Age of Surveillance Capitalism*, predictive algorithms represent a new form of “knowledge power” that operates through persistent data gathering and behavior alteration (Zuboff 93). Student-attendance-monitoring software, patient-risk-assessment healthcare algorithms, and employee-screening software for candidate suitability all demonstrate how AI has fulfilled the knowledge-production role that Foucault recognized in traditional institutions. These systems create what Foucault called “dividing practices” - processes that split and divide individuals on the grounds of institutional standards for normalcy (Foucault, *Power* 208).

The most direct application of Foucault’s paradigm is quite likely in AI systems’ surveillance function that operationalizes his panoptic principle. Reading Jeremy Bentham’s eighteenth-century plan for the Panopticon prison, Foucault showed how constant

visibility enables self-disciplined subjects (Foucault, *Discipline* 201). Contemporary algorithmic surveillance systems - from China's Social Credit System to social media monitoring - produce an on-screen panopticon where citizens self-reform in anticipation of judgment by algorithms (Browne 18). Such systems actualize what Foucault wrote about as "biopower" - ruling populations by statistical analysis and risk assessment (Foucault, *History of Sexuality* 139). The predictive policing algorithms decried by Eubanks, for instance, insinuate police surveillance in marginalized populations for all but purely objective data analysis, showing how AI reproduces existing relations of power in seeming to be technically neutral (Eubanks 76). This digital disciplining, like its institutional counterparts, produces knowledge that confirms its power, constituting a self-reinforcing cycle that is facilitated by Foucault's paradigm to question and challenge.

Truth Production and Epistemic Authority

Modern AI systems increasingly play assertive roles in determining truth that constitute epistemic hierarchies through their design and use. Search engines such as Google and large language models such as ChatGPT play roles that could perhaps be taken to represent Foucault's notion of "regimes of truth" - institutional mechanisms that constitute valid knowledge in society (Foucault, *Power/Knowledge* 131). Such technologies practice truth claims through computations such as ranking, sorting, and producing

information, often stating their outputs as authoritative but neutral. As Noble illustrates in *Algorithms of Oppression*, Google's algorithms consistently prefer some voices over others, effectively "determining what is knowable" through technical mediation (Noble 36). Such machine truth production is raising fundamental questions regarding control over power in digital knowledge economies, particularly when corporations own epistemic validation infrastructure.

The gate-keeping function of AI systems generates what Couldry and Hepp term "deep mediatization" - an effect in which algorithmic processes shape social reality by governing access to the public space of discourse (Couldry and Hepp 54). As ChatGPT generates answers or Google auto-completes keywords, it's doing that which Foucault termed the "ordering of things" - classifying and organizing knowledge in terms of structures in concordance with immanent logics of their data training and design parameters (Foucault, *Order of Things* xx). Scholarship on Wikipedia-based AI systems reveals that they reproduce the encyclopedia's systematic bias, over-representing Western attitudes to knowledge and reducing indigenous knowledge systems (Bender et al. 615). This algorithmic epistemicide - systematic elimination of particular knowledge traditions - echoes historical currents that Foucault wrote about in *The Archaeology of Knowledge*, whereby dominant discourses silence

other epistemologies through processes at play within institutions in terms of legitimation (Foucault 183).

These epistemic mechanisms have far-reaching democratic consequences. As Gillespie contends in *Custodians of the Internet*, algorithms on platforms function as "epistemic brokers" that settle on who speaks in digital publics (Gillespie 19). The case of Google demoting feminist blogs or ChatGPT not creating some political arguments but doing so quite willingly for others illustrates to us what Foucault termed the "polymorphous techniques of power" - indirect techniques that police discourse without overt censorship (Foucault, *History of Sexuality* 11). This epistemic rule, through algorithms, creates Rieder's "calculated publics" - publics constructed by and ruled by systems of algorithmic knowledge that naturalize certain worldviews (Rieder 34). Formulating these processes in terms of Foucault illustrates to us how AI systems have turned into modern truth apparatuses (*dispositifs*), technical formations that create and stabilize certain regimes of truth but express them in terms of objectivity and obligation.

Normalization & Social Control in the Age of Algorithms

Technological systems have developed artificial intelligence systems as powerful agents for social normalization, materializing Foucault's biopower

within data regimes. Productivity algorithms for workplace behavior like those at Amazon demonstrate this process by creating quantifiable measures for "ideal" employee behavior that uniformly identify and weed out variations from such artificially produced standards (Moore and Robinson 2987). Such systems take Foucault's account of disciplinary institutions to virtual spaces, where algorithmic tracking creates what he referred to as "docile bodies" - individuals who internalize and incorporate externally derived standards (Foucault, *Discipline and Punish* 138). Emotion recognition systems also represent this process of normalization in that they claim authoritative knowledge about affective states while imposing Western-centric models for emotional taxonomies over pluralistic cultural pasts (Stark 12). This management by algorithms of human behavior is an advanced version of what Foucault found to constitute "microphysics of power" - diffuse systems of power that operate on infinitesimal scales in everyday life (Foucault, *Power/Knowledge* 39).

China's social credit system is perhaps the most expansive example of AI-based biopolitical management. By integrating data from financial transactions, social media, and even jaywalker infractions, the system achieves Foucault's prophetic dream of "a society of control" based on ongoing supervision and scoring of citizens (Deleuze 7). The biopower template provided by Foucault explains

such developments as part of a historical process in which states tend to manage populations based on statistical knowledge and acts of calculation (Foucault, *Security, Territory, Population* 67). The social credit system does not punish infractions but rather works to constitute subjectivities by enforcing self-regulation by threat of algorithmic exclusion - one can argue, in Foucauldian terms, about "technology of the self" in digital language (Foucault, *History of Sexuality* 27). Such systems are that which is called "measurable life," where algorithmic measurement remakes citizenship, creditworthiness, and social value (Cheney-Lippold 165).

This normalization extends from state surveillance to commercial platforms commodifying conformity in conduct. Social media rewarding for interactions results in "algorithmic habitus" - habituated behavior that has been constructed about platform rewards (Bucher 32). Psychological impacts from such systems resemble Foucault's account of panoptic discipline, in which the threat of surveillance alters behavior (Foucault, *Discipline and Punish* 201). Scholarship on Uber drivers demonstrates that algorithmic management results in self-disciplining workers internalizing platform metrics as self-worth indicators (Rosenblat and Stark 3765). These examples together demonstrate that AI systems have developed to become that which Foucault catalogued as "apparatuses of security" - mechanisms that regulate populations not by direct

application of force but by setting rules for permissible conduct and enforcing them (Foucault, *Security* 20). This rulemaking paradox in algorithms is that it is seemingly neutral-technical but reproduces existing social hierarchies and enhances them, making Benjamin's "race-critical technology studies" crucial to lay open their exclusionary logics (Benjamin 72).

Resistance & Counter-Discourse in Algorithmic Societies

Foucault's dictum that "where there is power, there is resistance" (*History of Sexuality* 95) finds strong realization in new processes of resistance to AI systems. Such resistance manifests in terms of that which Foucault wrote about as "counter-conducts" - practices that disrupt regimes of rule and dominant truths (*Security, Territory, Population* 201). In an era of algorithms, these counter-conducts manifest in three ways: grass-roots resistance to unjust systems, efforts to open AI infrastructure to more democracy using technical intervention, and institutional efforts to regulate AI-based decision-making. Together, they constitute that which Hoffmann terms an "algorithmic accountability movement" that seeks to open to public sight and reform relationships of power inscribed in AI systems (Hoffmann 4).

Algorithmic activism has been an influential force for resistance that enacts Foucault's "insurrection of subjugated knowledges" to challenge dominant

technical discourses (Power/Knowledge, 81). The campaigns by the Algorithmic Justice League to challenge racially discriminatory facial recognition systems offer an example here, working through technical audit and discursive rhetoric to demonstrate how automatic systems amplify historical patterns of discrimination (Buolamwini and Gebru 1). In an equally compelling example, BantheScan effectively shamed multiple cities to outlaw police facial recognition, an example of Foucault's 'tactical polyvalence of discourses' - the ability for subordinated groups to reappropriate hegemonic universals to subvert them towards liberation (History of Sexuality 100). Such activist campaigns form an enactment of Benjamin's 'discriminatory design audits,' an application of an abolitionist approach to render legible how AI systems encode racially-based hierarchies (Benjamin 148).

Technical communities have built decentralized AI systems that actualize Foucault's vision for "heterotopias" - spaces that operate by other rules than governing institutions (Of Other Spaces). Federation-powered learning systems, for instance, remap computing while preserving data sovereignty (Yang et al. 18), and open-source AI efforts like EleutherAI subvert commercial control over massive language models (Biderman et al. 3). Such efforts actualize Eubanks's idea for "collective digital self-defense" in opposition to surveillance capitalism (Eubanks 189), creating what Foucault might see as

"technologies of the self" for an era of algorithms (History of Sexuality 26). The creation of explainable AI (XAI) and impact analyses for algorithms is another response, attempting to make obscure mechanisms of power legible and answerable - an analogy in computation to Foucault's plea to "reveal the political character of technical apparatuses" (Power/Knowledge 132).

Institutional control like that of the EU's GDPR and upcoming AI Act is that which Foucault would analyze as juridical internalization of resistance (*Discipline and Punish* 222). Article 22 in GDPR establishes an 'explainer's right' to automated decision-making tries to ward off algorithmic opacity in this manner through law (Wachter et al. 84). AI audit processes developed by researchers like Raji et al. perform Foucault's genealogical practice by tracing in which ways power is exercised through tech systems (Raji et al. 620). Such regulation-based counter-conducts, despite being problematic, demonstrate what Zuboff calls "the first groping steps toward a division in society that might work to protect and foster epistemic conditions for human freedom" (Zuboff 481). Such regulations create spaces for freedom within otherwise disciplinary regimes for Foucault (Ethics 284), meaning that resistance to algorithmic power is as much in institutional reform as in subversion in tech.

Conclusion

This discussion has revealed that artificial intelligence systems don't reflect relations of power in existence but, by their knowledge-producing capability, instead institute new regimes of truthmaking and rule. As warned by Foucault, "power produces knowledge. power and knowledge directly imply one another" (*Discipline and Punish* 27). From workplace monitor algorithms to social credits, we have seen that AI applies Foucault's concepts on biopower, disciplinary institutions, and governmentality to unprecedented breadth and complexity. Such systems create what one could perhaps term 'data colonialism' - an emerging epistemic mode of domination appropriating knowledge from populations to fuel algorithmic governance (Couldry and Mejias 336). Normalizing such conduct by cloaking it in rhetoric on being technically neutral is that which Foucault characterized as the "regime of truth's" most insidious side: concealing from sight their own historicity and political contingency (*Power/Knowledge* 133).

The future demands nothing short of a fundamental shift towards AI development based on openness, accountability, and participatory democracy. This entails centering marginalized communities in critique, as Costanza-Chock does in *Design Justice*, and in re-imagining algorithmic systems (Costanza-Chock 8). Technical solutions like explainable AI must be complemented by institutional ones like algorithmic impact assessments, tight public regulation, and

Zuboff's 'human sovereignty clauses' to ward off automated decision-making in ways of life (Zuboff 495). The European Union's proposed AI Act is an imperfect step in this direction, although such regulation is likely to reflect very much the very imbalances of power that it promises to address in the absence of accompanying grassroots mobilizations (Cath et al. 156). These efforts must be led by what Foucault wrote about as 'a new economy of power relations,' deflecting epistemic authority from technical elites (History of Sexuality 90).

Finally, this Foucauldian analysis leaves us with an intriguing challenge: Can AI's knowledge-power machine be reconceived as an instrument for emancipatory knowledge, not control? Feminist AI scholarship, such as that by Keyes, for example, shows us ways in which machine learning algorithms might be designed to "amplify rather than silence marginalized voices" (Keyes 14). Likewise, Indigenous AI efforts offer us ways in which technological innovation might be driven by non-dominant knowledge systems if the 'regime of truth' is turned productively on its side (Lewis et al. 5). As we learn from Foucault, power is not one but multiple, a 'site of multiple force relations' that can be remade. The challenge now is to create Benjamin's 'subversive technologies' that subvert AI's disciplinary logic, remaking it from an instrument of control to a means to an end to democratize knowledge production. In

this necessary task, Foucault's toolkit is still indispensable - not as an instruction manual for despair but as tools for unmasking and rewriting the relations of power beneath our algorithmic society.

Works Cited

- Beer, David. *The Social Power of Algorithms*. SAGE Publications, 2017.
- Bender, Emily M., et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 610-623.
- Benjamin, Ruha. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity, 2019.
- Biderman, Stella, et al. "Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task." *arXiv*, 2022, arXiv:2210.13382.
- Browne, Simone. *Dark Matters: On the Surveillance of Blackness*. Duke UP, 2015.
- Bucher, Taina. *If...Then: Algorithmic Power and Politics*. Oxford UP, 2018.
- Buolamwini, Joy, and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings of Machine Learning Research*, vol. 81, 2018, pp. 1-15.
- Cath, Corinne, et al. "Artificial Intelligence and the 'Good Society'." *AI & Society*, vol. 33, 2018, pp. 153-162.

- Cheney-Lippold, John. *We Are Data: Algorithms and the Making of Our Digital Selves*. NYU Press, 2017.
- Costanza-Chock, Sasha. *Design Justice: Community-Led Practices to Build the Worlds We Need*. MIT Press, 2020.
- Couldry, Nick, and Andreas Hepp. *The Mediated Construction of Reality*. Polity, 2017.
- Deleuze, Gilles. "Postscript on the Societies of Control." *October*, vol. 59, 1992, pp. 3-7.
- Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, 2018.
- Foucault, Michel. "Of Other Spaces." *Diacritics*, vol. 16, no. 1, 1986, pp. 22-27.
- Foucault, Michel. *Power/Knowledge: Selected Interviews and Other Writings 1972-1977*. Edited by Colin Gordon, Pantheon, 1980.
- Foucault, Michel. *Security, Territory, Population: Lectures at the Collège de France 1977-1978*. Edited by Michel Senellart, Picador, 2007.
- Foucault, Michel. *Discipline and Punish: The Birth of the Prison*. Translated by Alan Sheridan, Vintage, 1995.
- Foucault, Michel. *Ethics: Subjectivity and Truth*. Edited by Paul Rabinow, New Press, 1997.
- Foucault, Michel. *The Archaeology of Knowledge*. Translated by A.M. Sheridan Smith, Pantheon, 1972.

- Foucault, Michel. *The History of Sexuality, Volume 1: An Introduction*. Translated by Robert Hurley, Vintage, 1990.
- Foucault, Michel. *The Order of Things: An Archaeology of the Human Sciences*. Vintage, 1994.
- Gillespie, Tarleton. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale UP, 2018.
- Hoffmann, Anna Lauren. "Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse." *Information, Communication & Society*, vol. 22, no. 7, 2019, pp. 900-915.
- Introna, Lucas D. "Algorithms, Governance, and Governmentality." *Science, Technology, & Human Values*, vol. 41, no. 1, 2016, pp. 17-49.
- Keyes, Os. "The Misgendering Machines." *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, 2018, pp. 1-22.
- Lewis, Jason Edward, et al. "Indigenous Protocol and Artificial Intelligence Position Paper." *Indigenous AI*, 2020.
- Moore, Phoebe, and Andrew Robinson. "The Quantified Self: What Counts in the Neoliberal Workplace." *New Media & Society*, vol. 18, no. 11, 2016, pp. 2774-2792.

- Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018.
- Raji, Inioluwa Deborah, et al. "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing." *FACCT '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 33-44.
- Rieder, Bernhard. "Big Data and the Paradox of Diversity." *Digital Culture & Society*, vol. 2, no. 2, 2016, pp. 21-40.
- Rosenblat, Alex, and Luke Stark. "Algorithmic Labor and Information Asymmetries: A Case Study of Uber's Drivers." *International Journal of Communication*, vol. 10, 2016, pp. 3758-3784.
- Stark, Luke. "Facial Recognition, Emotion and Race in Animated Social Media." *First Monday*, vol. 23, no. 9, 2018.
- Wachter, Sandra, et al. "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation." *International Data Privacy Law*, vol. 7, no. 2, 2017, pp. 76-99.
- Yang, Qiang, et al. "Federated Learning." *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 13, no. 3, 2019, pp. 1-207.
- Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs, 2019.

