CONCEPTUAL RESEARCH ARTICLE

Human Agency in the Age of Artificial Intelligence

From Controlling Technical Systems to Shaping Societal Capacity for Action

A theoretical contribution and research agenda

Working paper · July 2026

Abstract

Current discourse on artificial intelligence (AI) is often shaped by two limitations. First, AI is framed as a technological competitor to human agency. Second, governance is treated primarily as the control of technical risks. This conceptual article proposes a different perspective. Its central question is not whether human or machine agency increases in the abstract, but how AI redistributes, expands, constrains, and transforms the conditions of human agency. Drawing on the capability approach, social-cognitive theories of agency, structuration theory, self-determination theory, research on work autonomy, and adaptive governance, the article develops a five-dimensional model. Capacity comprises knowledge, skills, and means; Room denotes institutionally, materially, and socially available space for action; Will refers to motivational and volitional activation; Direction concerns the epistemic and normative formation of goals; and Identity captures the self- and role-conceptions from which responsibilities and purposes emerge. The model treats agency as a relational and dynamic configuration rather than an additive personal attribute. Building on it, the article proposes an Agency Impact Assessment that combines distributional analysis, uncertainty mapping, reversible experimentation, and continuous learning. Applications to knowledge work, algorithmic management, education, social media, and political economy illustrate the framework. The article argues that AI governance should be complemented by the deliberate societal governance of the development and distribution of human agency. It concludes with limitations, operationalization requirements, and an empirical research agenda.

Keywords: human agency; artificial intelligence; capability approach; autonomy; algorithmic management; adaptive governance; political economy

1. Introduction

Generative AI shifts the boundary of what individuals and organizations can produce with limited expertise, time, and capital. Text, software, analysis, and media can be created more rapidly. At the same time, recommender systems, automated decisions, and algorithmic management can structure or narrow human discretion. This coexistence of enablement and constraint is poorly captured by debates that treat AI either as a neutral tool or as an autonomous rival.

The prevailing control question—“How can we control AI?”—is indispensable when safety, reliability, and accountability are at stake. Yet it can obscure the fact that technical systems are embedded in property regimes, employment relations, legal rules, platform markets, and cultural interpretations. The socially consequential variable is therefore not model capability alone, but who can act, decide, object, learn, and pursue a conception of the future through its deployment.

This article develops human agency as a normative and analytical reference point for AI-rich societies. It makes three contributions. First, it synthesizes dispersed theoretical traditions into a five-dimensional model. Second, it treats AI effects as questions of distribution and power: the same application may increase workers’ Capacity while reducing their Room through surveillance. Third, it translates the theory into an adaptive governance procedure designed to remain capable of learning under uncertainty.

The article is a conceptual theory-building contribution, not a systematic review or an empirical validation study. Its constructs are proposed as testable distinctions. Their scientific value depends on subsequent operationalization, contextual comparison, and potential falsification.

2. Literature Review

2.1 Agency between causal efficacy and social structure

At a minimal level, agency denotes the ability to produce effects through action. Bandura (2001, 2006) emphasizes intentionality, forethought, self-reactiveness, and self-reflectiveness. His account of collective agency extends analysis beyond isolated individuals. Giddens (1984) defines agency through the possibility of acting otherwise and relates it to the duality of structure: rules and resources enable action and are reproduced through action. Archer (2003) instead foregrounds reflexivity as the mediation between structural conditions and personal concerns. Emirbayer and Mische (1998) reveal the temporal architecture of agency, in which habitual patterns, projective imagination, and practical evaluation interact.

These approaches caution against treating agency as a stable internal possession. Agency is relational: it arises through the interaction of actor, situation, resources, expectations, and temporal orientation. This is precisely what makes the concept valuable for the study of AI. A language model may make task execution easier while organizational pressure, absent appeal mechanisms, or epistemic dependence diminishes effective human control.

2.2 The capability approach, freedom, and inequality

Sen's capability approach shifts the evaluation of welfare away from resources or utility toward substantive opportunities: what people are genuinely able to do and to be (Sen, 1999). Nussbaum (2011) specifies central capabilities and their political thresholds. The approach is highly relevant to AI because identical technical resources do not generate identical opportunities. Education, time, income, infrastructure, legal status, and social recognition shape whether access to AI can be converted into effective action.

The present model uses Capacity more narrowly than Sen uses capability: Capacity primarily refers to individual and collectively available abilities and means. Room captures conversion conditions, freedoms, and institutional opportunities. This separation helps distinguish interventions that strengthen competence from those that merely provide formal access. It also creates conceptual overlap that must be tested empirically against established capability constructs.

2.3 Motivation, autonomy, identity, and orientation

Self-determination theory links autonomous motivation to basic needs for autonomy, competence, and relatedness (Deci & Ryan, 2000; Ryan & Deci, 2017). Self-efficacy shapes whether people undertake

challenges and persist after setbacks (Bandura, 1997). Research on work design similarly identifies autonomy, task significance, and feedback as central to psychological states and motivation (Hackman & Oldham, 1976). Psychological safety expands the perceived scope for questions, disagreement, and learning from error (Edmondson, 1999).

Will is therefore not used here as a strict metaphysical category in the Schopenhauerian sense, but as volitional activation: the readiness to translate a direction into action despite effort, uncertainty, and resistance. Direction denotes the formation and selection of desired future states. Identity denotes relatively durable self- and role-conceptions that make particular goals, duties, and possibilities intelligible. Identity is plural and revisable; it should not be mistaken for an immutable essence.

2.4 Technology, power, and algorithmic management

Technology is political insofar as artifacts and infrastructures can stabilize social arrangements and distribute possibilities for action unequally (Winner, 1980). Foucault's account of productive power directs attention to the ways subjects are not only constrained but constituted through norms, measurement, and self-surveillance (Foucault, 1977). Digital platforms intensify these dynamics through data extraction, behavioral prediction, and asymmetric visibility (Zuboff, 2019; Crawford, 2021).

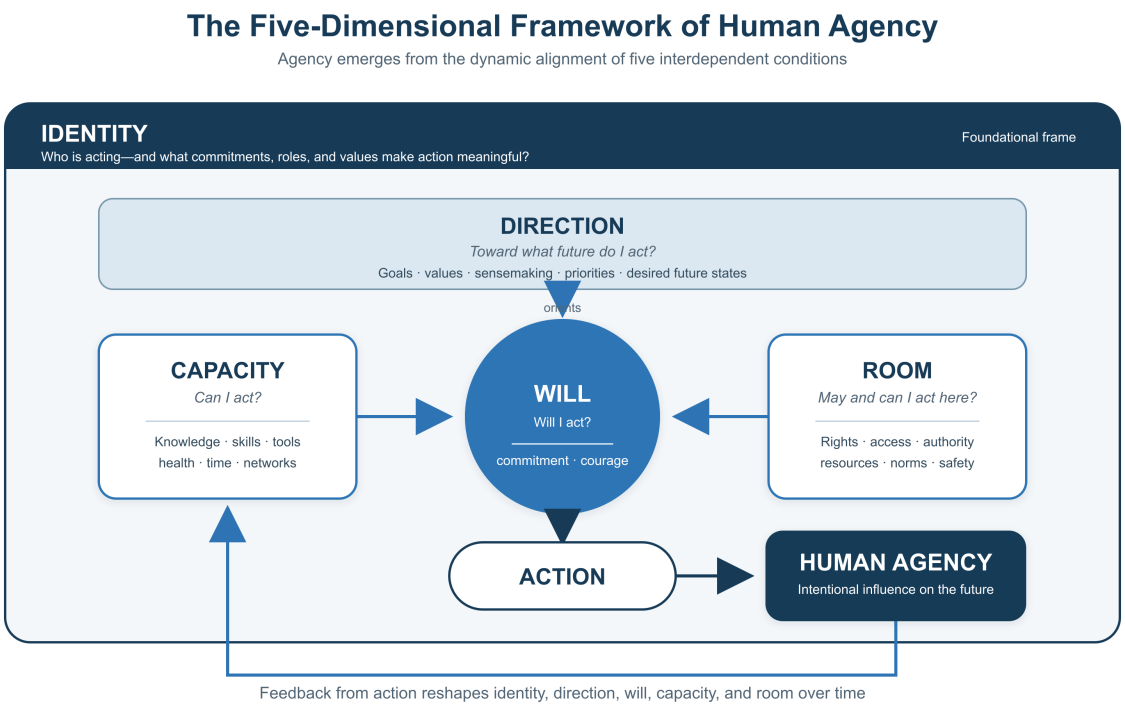
Research on algorithmic management accordingly reveals a duality. Algorithms may provide information, facilitate coordination, and enable autonomy; they may also intensify work, obscure decisions, and strengthen managerial control (Kellogg et al., 2020; Meijerink & Bondarouk, 2023). The automation–augmentation paradox suggests that the same technology may substitute for or complement human work depending on task architecture and organizational design (Raisch & Krakowski, 2021). AI effects are therefore neither technologically predetermined nor arbitrary. They are mediated by design, power, and institutional feedback.

2.5 From AI ethics to adaptive governance

Existing governance frameworks provide essential foundations. UNESCO's Recommendation connects human rights, human oversight, accountability, education, and adaptive multi-stakeholder governance (UNESCO, 2021). The NIST AI Risk Management Framework organizes continuing practice through Govern, Map, Measure, and Manage, while emphasizing sociotechnical context and clearly defined human roles (NIST, 2023). Floridi et al. (2018) articulate principles for a "good AI society," while Ostrom (2010) demonstrates how polycentric governance can enable learning across multiple centers of decision-making.

Agency-centered governance does not replace safety, rights, or risk frameworks. It adds a positive and distributional question: Which people and groups gain or lose the long-term ability to shape their own and shared futures?

3. A Five-Dimensional Framework of Human Agency



Working definition: Human agency is the socially embedded ability of persons or collectives to perceive and evaluate possibilities, select a direction, commit to it, and act effectively upon future states. Agency therefore combines ability, freedom, activation, orientation, and self-relation.

Dimension	Guiding question	Core components	Typical AI effects
Capacity	Can I act?	Knowledge, skills, health, tools, time, networks	Augmentation, learning, deskilling, dependence
Room	May and can I act here?	Rights, resources, authority, access, norms, psychological safety	Access, surveillance, standardization, appeal rights
Will	Will I act?	Motivation, self-efficacy, courage, persistence, ownership	Relief, activation, passivity, automation bias
Direction	Toward what do I act?	Goals, values, sensemaking, priorities, images of the future	Expanded options, nudging, attention steering, goal substitution
Identity	Who acts, and for what?	Self-concept, roles, belonging, authorship, commitments	Role change, identity threat, new professional identities

Table 1. Dimensions of the proposed agency model. Author's conceptualization.

3.1 Relational and non-additive structure

The dimensions are analytically distinguishable but causally interdependent. Identity shapes which directions appear fitting or obligatory; Direction focuses Will; Will mobilizes Capacity; and Room helps determine whether mobilized abilities can become effective. Feedback also runs in reverse. Successful action can strengthen self-efficacy and identity, while repeated obstruction can erode motivation and narrow future horizons.

A multiplicative metaphor—$A \approx C \times R \times W$, framed by Direction and Identity—illustrates bottlenecks, but it is not a validated equation. True zero values are rare, dimensions may partly compensate for one another, and power operates across several dimensions at once. Agency is better understood as a configuration: it arises when the dimensions are sufficiently aligned for a particular actor, action, context, and time.

3.2 Levels and distribution

Agency exists at individual, collective, and institutional levels. Teams may possess agency through shared knowledge, decision rights, and common aims even where each member contributes only partially. Institutions shape meta-agency by determining who may revise rules, inspect data, select models, or challenge decisions. The relevant outcome is therefore not only average agency, but its distribution across income, education, gender, disability, employment status, geography, and organizational position.

3.3 Power as influence over the agency of others

The model implies a relational view of power: power includes the ability to open, channel, or constrain the agency of others. AI providers affect Capacity through access to models, Room through terms of service, Direction through interface defaults, and Identity through the categories in which users are addressed. Employers determine whether productivity gains become learning time, reduced hours, higher pay, or intensified monitoring. AI does not redistribute agency by itself; institutions redistribute it through AI.

4. Agency-Centered Analysis of Selected Domains

4.1 Knowledge work and AI assistants

AI assistants can substantially increase Capacity by supplying drafts, translations, code, and syntheses. Short-term performance, however, is not equivalent to durable competence. If outputs are accepted without evaluation, practice, error diagnosis, and domain understanding may weaken. Good design therefore separates assistance from the transfer of responsibility: it exposes uncertainty, requires reasoned selection, and preserves opportunities for independent thought. What matters is not simply

whether a human remains “in the loop,” but whether that person possesses the knowledge, time, authority, and genuine possibility to object.

4.2 Work, leadership, and psychological safety

From an agency perspective, leadership designs the conditions under which others act. Micromanagement narrows Room, can weaken Will, and recasts professional Identity from responsible contributor to executor. Empowering leadership expands discretion but must also provide resources, capability development, and clear boundaries of responsibility; otherwise autonomy becomes responsibility without power. Psychological safety expands perceived Room for dissent and learning from error, but cannot substitute for job security or formal rights of appeal.

4.3 Algorithmic management

Agency distribution is especially visible in algorithmic allocation and performance evaluation. Management gains Capacity for observation and optimization, while workers may lose Room when targets, pace, and ratings cannot be negotiated. Transparency alone is insufficient. Agency requires intelligibility, contestability, collective participation, alternative decision channels, and protection from retaliation. An explainable system without an effective right to depart from its recommendation remains a system of control.

4.4 Education

AI tutors may broaden access to feedback and individualized support. If they are optimized primarily for correct answers, however, they may displace productive struggle, epistemic independence, and the development of Direction. Agency-centered education therefore measures not only learning outcomes but also transfer, judgment, self-efficacy, goal formation, and the capacity to reject an AI output critically. Educators require Room to conduct local experiments rather than merely implement centralized platform rules.

4.5 Social media and the public sphere

Social media expand Capacity and Room for publication, networking, and organization. At the same time, recommendation and reward systems shape attention, social recognition, and thematic priorities. Will may be redirected into short-term reaction, Direction fragmented, and Identity tied performatively to metrics. The normative question is not only whether content is harmful, but whether the architecture supports sustained, reflective, and collective capacity for action.

4.6 Wages, inequality, and political economy

Income is not merely a resource for consumption; it amplifies agency. It purchases time, education, mobility, legal support, technical access, and the ability to take risks. If AI-driven productivity gains accrue predominantly to the owners of models, data, and capital, aggregate Capacity may rise while the

Room and Will of broad groups contract. Precarity shortens planning horizons and binds attention to immediate security. Wage and ownership questions are therefore questions of mass agency. Relevant policies include competition, worker participation, portable rights, social protection, access to public digital infrastructure, and a distribution of productivity gains that creates time and opportunities for learning.

5. Adaptive Agency Governance

Because the long-term effects of generative AI are uncertain, context-dependent, and recursive, neither a technological master plan nor passive waiting is adequate. This article proposes an Agency Impact Assessment (AIA) to complement existing risk and fundamental-rights assessments. It combines normative clarification, systems mapping, hypothesis testing, and mandatory revision.

1. **Clarify purpose and boundaries.** Define legitimate aims, non-negotiable rights, accountable bodies, and stopping rules before deployment.
2. **Map agency.** Establish a baseline and expected changes in Capacity, Room, Will, Direction, and Identity for each affected group.
3. **Analyze distribution and power.** Identify winners, losers, decision points, data access, ownership, appeal mechanisms, and cumulative effects.
4. **Classify knowledge.** Separate evidence, plausible hypotheses, and genuine unknowns; document assumptions and expected observations.
5. **Run safe-to-fail trials.** Use small, time-limited, reversible pilots with safeguards, comparison conditions, and heterogeneous participants.
6. **Measure multiple dimensions.** Track competence, autonomy, motivation, goal clarity, identity, workload, performance, and distributional outcomes.
7. **Decide and revise.** Scale, modify, or stop; reassess consequences regularly and retain genuine exit options.

5.1 Measurement logic

An AIA should combine qualitative and quantitative evidence. Capacity can be assessed through task performance, transfer, and error sensitivity; Room through discretion, resources, contestability, and psychological safety; Will through autonomous motivation, initiative, and self-efficacy; Direction through goal clarity, value congruence, and prioritization; and Identity through professional authorship, role coherence, and experienced responsibility. Objective institutional indicators are also required, because reported autonomy can remain high even where people have adapted to constraint.

5.2 Decision rules under uncertainty

Reversibility is not a license for unrestricted experimentation. High-risk applications require strict legal and ethical limits in advance. Within permissible domains, decisions should be hypothesis-led. A deployment should be judged agency-enhancing only where gains in one dimension are not purchased through severe and uncompensated losses in another dimension or group. Where trade-offs are unavoidable, they must be decided transparently as political choices rather than concealed as technical optimization.

6. Discussion

6.1 From human–machine competition to institutional configuration

The frame “human agency versus AI agency” captures real instances of delegated decision-making, but remains too coarse. An AI system may display functional autonomy within a process without possessing property rights, legitimate purposes, or social authority. The decisive questions are who develops it, who may deploy it, who can audit or stop it, and who captures its benefits. The appropriate unit of analysis is therefore the human–AI–institution configuration.

This does not diminish technical safety. It connects safety with political economy. A reliable system may concentrate agency; an unreliable one may destroy Room through decisions at scale. Alignment with individual preferences is also insufficient where systems shape those preferences or exchange long-term judgment for short-term convenience.

6.2 Fear of the unfamiliar and warranted concern

Public reactions to AI may be influenced by uncertainty avoidance, anthropomorphism, status threat, and the construction of AI as an unfamiliar actor. A direct analogy to racism would nevertheless be categorically and normatively misleading: AI systems are manufactured artifacts, not historically oppressed human groups, and are not straightforwardly moral subjects. The more defensible hypothesis concerns partially shared psychological mechanisms of unfamiliarity and threat perception.

Agency analysis separates psychological response from institutional evidence. Fear may be excessive where an application genuinely enables people; acceptance may be misguided where convenience quietly undermines agency. This distinction supports criticism of both indiscriminate technological pessimism and uncritical solutionism.

6.3 Relationship to existing approaches

The framework is closest to the capability approach, while differentiating motivational activation, orientation, and identity more explicitly. Relative to Responsible AI, it shifts attention from system characteristics to distributed human possibilities for action. Relative to Human-Centered AI, it insists

that “the human” is not a unitary stakeholder. Relative to conventional autonomy research, it adds resources, identity, and social distribution. Its contribution therefore lies less in entirely new components than in their integrated diagnostic use and translation into adaptive governance.

7. Implications

7.1 Organizations

Organizations should evaluate AI deployments by more than productivity and user acceptance. Procurement, process design, and leadership must include learning time, decision rights, channels for objection, and the distribution of efficiency gains. Worker representatives and directly affected groups should participate in goal definition, pilot design, and evaluation. Roles should align responsibility with knowledge and authority.

7.2 Education

Curricula should combine AI literacy with epistemic agency: framing problems, checking sources, tolerating uncertainty, justifying goals, and accepting responsibility for outcomes. Assessment must distinguish improved output from durable judgment. Educational institutions should also address changes in professional authorship and identity explicitly.

7.3 Public policy and regulation

Policy should complement rights and risk assessments with distributional indicators of agency. Collective bargaining power, effective appeal against automated decisions, access to data and models, public digital infrastructure, competition policy, and social protection are especially important. Agency cannot replace prosperity, justice, or rights as a single master metric. It is a cross-cutting perspective that reveals when people are formally protected but practically unable to act.

8. Limitations

First, the framework contains conceptual overlaps: Capacity and Room intersect with capabilities and resources; Will with motivation and self-efficacy; and Direction with Identity. Discriminant validity remains to be demonstrated. Second, a strong agency norm may undervalue care, dependence, and legitimate non-action. People need not maximize action continuously; the freedom to delegate or join collective decisions can itself express agency. Third, agency is not inherently good. Powerful actors may use high agency for harmful ends, so the framework requires external constraints grounded in rights, justice, and the prevention of harm.

Fourth, measurement is normatively charged: those who define indicators exercise power. Fifth, the dimensions may vary culturally and institutionally; individualistic notions of autonomy must not displace

collective and relational forms. Sixth, this article offers a narrative synthesis. It cannot claim completeness or causal validity without systematic review and empirical research.

9. Future Research

A research program should begin with qualitative interviews across domains to refine the five dimensions from the perspective of affected people. Item development, factor analysis, and tests of convergent and discriminant validity should follow. Longitudinal studies should distinguish short-term productivity from trajectories of competence and motivation. Field experiments could compare governance designs, including voluntary versus mandatory use, explainable recommendations versus contestable decisions, and individual access versus team-shared AI.

Distributional research is particularly urgent. Who receives additional time, income, and decision power through AI, and who becomes more intensively measured and standardized? Multilevel models should connect individual agency, team climate, organizational rules, and market structure. Comparative work should examine democratic, authoritarian, and polycentric governance arrangements, as well as the conditions under which collective agency complements or crowds out individual agency.

10. Conclusion

The central social question of the AI age is not only what machines can do or how they can be controlled. It is which people and collectives, under which institutional conditions, can effectively shape their futures. The five-dimensional model shows that technological enablement may coexist with institutional restriction, motivational passivity, externally imposed direction, or identity loss. Conversely, well-designed AI may expand skills, discretion, and collective learning capacity.

Agency-centered governance therefore joins technical risk, power, and adaptive learning. Its goal is not the maximization of agency at any cost, but a just, responsible, and developmental distribution of human capacity for action within clear legal and ethical boundaries. Whether the proposed framework can serve that purpose is not a matter of conceptual elegance alone, but of future empirical performance.

References

Archer, M. S. (2003). Structure, agency and the internal conversation. Cambridge University Press.

Bandura, A. (1997). Self-efficacy: The exercise of control. W. H. Freeman.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. Annual Review of Psychology, 52, 1–26. https://doi.org/10.1146/annurev.psych.52.1.1

Bandura, A. (2006). Toward a psychology of human agency. Perspectives on Psychological Science, 1(2), 164–180. https://doi.org/10.1111/j.1745-6916.2006.00011.x

Beer, S. (1972). Brain of the firm. Allen Lane.

Crawford, K. (2021). Atlas of AI. Yale University Press.

Deci, E. L., & Ryan, R. M. (2000). The “what” and “why” of goal pursuits: Human needs and the self-determination of behavior. Psychological Inquiry, 11(4), 227–268. https://doi.org/10.1207/S15327965PLI1104_01

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. Administrative Science Quarterly, 44(2), 350–383. https://doi.org/10.2307/2666999

Emirbayer, M., & Mische, A. (1998). What is agency? American Journal of Sociology, 103(4), 962–1023. https://doi.org/10.1086/231294

Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People—An ethical framework for a good AI society. Minds and Machines, 28, 689–707. https://doi.org/10.1007/s11023-018-9482-5

Foucault, M. (1977). Discipline and punish: The birth of the prison (A. Sheridan, Trans.). Pantheon Books.

Giddens, A. (1984). The constitution of society. University of California Press.

Hackman, J. R., & Oldham, G. R. (1976). Motivation through the design of work: Test of a theory. Organizational Behavior and Human Performance, 16(2), 250–279. https://doi.org/10.1016/0030-5073(76)90016-7

Illich, I. (1973). Tools for conviviality. Harper & Row.

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. Academy of Management Annals, 14(1), 366–410. https://doi.org/10.5465/annals.2018.0174

Meijerink, J., & Bondarouk, T. (2023). The duality of algorithmic management: Toward a research agenda on HRM algorithms, autonomy and value creation. Human Resource Management Review, 33(1), 100876. https://doi.org/10.1016/j.hrmr.2021.100876

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1

Nussbaum, M. C. (2011). Creating capabilities: The human development approach. Harvard University Press.

Ostrom, E. (2010). Beyond markets and states: Polycentric governance of complex economic systems. American Economic Review, 100(3), 641–672. https://doi.org/10.1257/aer.100.3.641

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. Academy of Management Review, 46(1), 192–210. https://doi.org/10.5465/amr.2018.0072

Ryan, R. M., & Deci, E. L. (2017). Self-determination theory. Guilford Press.

Sen, A. (1999). Development as freedom. Oxford University Press.

UNESCO. (2021). Recommendation on the ethics of artificial intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137

Weick, K. E. (1995). Sensemaking in organizations. Sage.

Winner, L. (1980). Do artifacts have politics? Daedalus, 109(1), 121–136.

Zuboff, S. (2019). The age of surveillance capitalism. PublicAffairs.