Collective and Individual Decisions
e-ISSN: 3071-7973

The Optimization Trap: How Artificial Intelligence Diminishes Decision Rationality in Organizations

Dariusz Jemielniak[1]

Abstract

A common assumption holds that artificial intelligence improves organizational decisions because it processes more data, computes faster, and behaves more consistently than human decision-makers. This article argues the opposite for a large class of consequential decisions: AI frequently *decreases* decision rationality in organizations, even as it improves prediction accuracy, processing speed, and internal consistency. The problem is not a malfunction. It is that AI succeeds at something narrower than what organizations actually need, while making that narrow success resemble the whole picture. Drawing on Simon's account of bounded rationality and the behavioral tradition that followed, the paper distinguishes a rich conception of organizational rationality – one requiring contextual sensitivity, interpretive flexibility, value awareness, and reflexive judgment – from the thin, optimization-based notion that dominates managerial discourse on AI. It then identifies four interconnected mechanisms through which algorithmic systems erode the richer capacity: proxy optimization, which substitutes measurable targets for meaningful goals; automation bias, which compresses human judgment around algorithmic defaults; decontextualization, which strips cases of the particularity that good decisions depend on; and the suppression of reflexive judgment, which prevents organizations from questioning their own premises. These mechanisms produce cumulative organizational consequences: weakened accountability, diminished contestability, and reduced capacity for adaptive learning. The

[1] Dariusz Jemielniak – Kozminski University, Katedra Zarządzania W Społeczeństwie Sieciowym, Warsaw, Poland, e-mail: darekj@kozminski.edu.pl, https://orcid.org/0000-0002-3745-7931

argument is not that organizations should abandon AI, but that they should stop treating it as a rationality upgrade. AI is a capable tool for narrow computational tasks and a poor substitute for organizational judgment. The organizations most at risk are those that have forgotten the difference.

Keywords: artificial intelligence, organizational decision-making, bounded rationality, algorithmic management, proxy optimization, automation bias, accountability

Introduction

A widespread faith pervades contemporary organizational thinking: that artificial intelligence, by processing more data and computing more efficiently than humans, necessarily produces better decisions. This faith is understandable. AI systems are fast, consistent, tireless in operation, and impressively calculative. They do not tire in the human sense, and they often appear less erratic than human decision-makers, which makes it tempting to treat them as more rational by definition. The conclusion seems obvious – if bad decisions stem from cognitive limitations, then removing those limitations through algorithmic processing should yield better outcomes.

The conclusion is wrong. Or rather, it rests on an impoverished understanding of what rationality means in organizational life. This article argues that AI frequently decreases decision rationality in organizations, even when it improves prediction accuracy, processing speed, or internal consistency. The mechanism is not malfunction. AI does not reduce rationality by failing at what it does. It reduces rationality by succeeding at something narrower than what organizations actually need – and by making that narrow success look like the whole picture.

The argument proceeds in four steps. First, I distinguish a richer concept of organizational rationality from the thin, optimization-based notion that dominates AI discourse. Second, I explain why AI so easily passes for a rational technology. Third – and this is the core of the paper – I show how AI diminishes rationality through proxy optimization, automation bias, decontextualization, and the suppression of reflexive judgment. Finally, I trace the organizational consequences: weakened contestability, eroded accountability, and diminished capacity for adaptive learning.

A clarification at the outset. I do not argue that AI is inherently irrational or that organizations should abandon algorithmic tools. I argue something more specific and, I think, more troubling: that the very features that make AI appear rational – its formalism, consistency, and data-

dependence – are the same features that systematically erode the capacities organizations need most when facing ambiguity, contested values, and genuine uncertainty.

Rethinking rationality in organizational decision-making

The standard case for AI in organizations leans on a particular model of rationality: the rational actor who surveys options, estimates outcomes, and selects the alternative that maximizes expected utility. This is the classical model that Herbert Simon spent decades dismantling (Simon, 1955; Simon, 1997). Simon's insight was not merely that humans fall short of the classical ideal – everyone concedes that – but that the ideal itself mischaracterizes what good decision-making actually requires.

Bounded rationality, as Simon formulated it, is not a deficiency. It is the realistic cognitive condition under which all organizational decisions take place. Decision-makers operate with incomplete information, limited computational capacity, unclear preferences, and time constraints. They satisfice rather than optimize. They use heuristics that are often remarkably well adapted to their environments (Gigerenzer, & Brighton, 2009). This is not a bug waiting to be patched by better technology. It is the structure of organizational judgment.

March extended this analysis further, but the classic formulation of organizations as settings of problematic preferences, unclear technologies, and fluid participation comes from Cohen, March, and Olsen's garbage can model (Cohen, March, & Olsen 1972; March, 1994). Organizations frequently do not know what they want until after they have acted. Goals are discovered, revised, and contested through the process of deciding itself. This is not irrationality. It is the nature of complex institutional life.

A genuinely rational organizational decision, in this richer sense, requires several capacities that standard optimization cannot provide. It requires sensitivity to context – the ability to recognize when established categories no longer fit the situation. It requires interpretive flexibility – the capacity to reframe problems, not just solve them as given. It requires value awareness, meaning the willingness to ask whether the objective being pursued is worth pursuing at all. And it requires what Dreyfus (1992) called practical expertise: the ability to perceive relevant features of a situation without reducing them to explicit rules.

None of this makes rationality mysterious or anti-scientific. It makes it richer than a maximization function. The question for AI in organizations is therefore not whether it calculates

better than humans. It almost certainly does. The question is whether better calculation translates into more rational decisions. My claim is that, in many consequential cases, it does not.

Why AI appears rational

Before showing how AI diminishes rationality, it is worth understanding why its rationality is so widely assumed. AI possesses several properties that closely mimic what people associate with rational deliberation.

Consistency, first. An AI system applies the same decision rule every time, without variation. It does not have good days and bad days. For organizations under institutional pressure to demonstrate fairness and procedural regularity, this property is enormously attractive. Consistency, however, is not the same as correctness. A system can be perfectly consistent while applying a rule that is poorly specified, contextually inappropriate, or based on biased training data (O'Neil, 2016). Consistency means the system repeats its logic. It does not mean the logic is sound.

Speed and scale, second. AI can process information orders of magnitude faster than any human decision-maker, and across datasets no individual could survey. This capacity generates an impression of comprehensiveness – the system has 'seen everything' – that is largely illusory. What the system has done is computed over a dataset that someone selected, cleaned, and formatted according to prior choices about what counts as relevant. The speed is real. The comprehensiveness is a constructed artifact.

Third, formalization. AI outputs arrive in the language of precision: scores, probabilities, classifications, rankings. This formalism carries an aura of objectivity. A hiring algorithm does not say, 'I have a gut feeling about this candidate.' It says the candidate scores 73 out of 100. The number suggests rigor and impartiality, even when the underlying model reflects historical biases, contested assumptions, or arbitrary threshold decisions (Raghavan et al., 2020). Formalization is not objectivity. It is the appearance of objectivity, and the two are dangerously easy to confuse.

Prediction, finally. Agrawal, Gans, and Goldfarb (2018) have argued that AI is fundamentally a prediction technology – it reduces the cost of prediction and thereby changes how organizations make decisions. This framing is useful but incomplete. Better prediction helps when the decision environment is stable, the relevant variables are known, and the objective function is clear. Many organizational decisions do not meet these conditions. Prediction under fixed assumptions does not help an organization that needs to reconsider its assumptions.

These four properties – consistency, scale, formalization, and prediction – combine to produce something that looks very much like rational superiority. The AI system appears to do what the rational actor model always prescribed: process all available information and select the optimal response. What this appearance conceals is that the system has not overcome bounded rationality. It has merely shifted the boundaries, replacing some human limitations with different, less visible, and often less tractable machine limitations.

Why AI decreases decision rationality

This section develops the central argument. I identify four interconnected mechanisms through which AI diminishes organizational rationality: proxy optimization, automation bias, decontextualization, and the suppression of reflexive judgment.

Proxy optimization and the measurement trap

Organizational goals are complicated things. 'Improve patient health outcomes.' 'Hire the best candidates.' 'Reduce recidivism.' These objectives sound clear until you try to operationalize them for an algorithm. At that point, you must translate ambiguous goals into measurable indicators, and measurable indicators into computable targets. The algorithm then optimizes against those targets as though they were the goals.

This is the logic of proxy optimization, and it is deeply corrosive. Campbell (1979) formulated the problem decades before AI: the more a quantitative indicator is used for decision-making, the more it distorts the process it is supposed to monitor. A related point is captured by Goodhart's observation that statistical regularities tend to break down once used for control purposes; in later discussion, this idea is often paraphrased as the claim that a measure ceases to be useful once it becomes a target (Goodhart, 1975; Strathern, 1997). This applies with particular force to algorithmic systems, because those systems optimize with a thoroughness that amplifies every distortion the proxy introduces.

Obermeyer et al. (2019) documented a striking case. A widely used healthcare algorithm assigned risk scores to patients to determine who would receive additional care. The algorithm used healthcare costs as a proxy for healthcare needs. Because Black patients historically received less spending for equivalent levels of illness, the algorithm systematically underestimated their needs. The proxy – cost – was technically measurable. It was also deeply misleading. The better

the algorithm optimized against this proxy, the more efficiently it reproduced the underlying inequity.

This is not a story about a broken algorithm. The algorithm worked exactly as designed. The problem was structural. The proxy captured something measurable but not what mattered. And the system's very efficiency prevented anyone from noticing the gap – because the outputs looked precise, consistent, and well calibrated.

Muller (2018) has traced how metric fixation – the belief that quantitative indicators are inherently more trustworthy than qualitative judgment – reshapes institutional behavior across domains. When AI enters this picture, metric fixation intensifies. The algorithm demands operationalized targets. Organizations supply them. The algorithm optimizes. And the distance between what is being optimized and what the organization actually needs grows wider, often without anyone noticing, because the numbers keep improving.

Espeland and Sauder (2007) showed how rankings restructure the behavior of the institutions being ranked, a phenomenon they called reactivity. AI-driven metrics produce the same dynamic but faster, at larger scale, and with less visibility. When an organization optimizes its AI systems against proxy metrics, it does not merely track performance. It reshapes what performance means.

Automation bias

Automation bias is the tendency of human decision-makers to defer to automated recommendations even when those recommendations are wrong, and even when contradictory evidence is available (Parasuraman, & Manzey, 2010). This is not laziness. It is a robust cognitive phenomenon, well documented in aviation, healthcare, and military contexts, and it intensifies under precisely the conditions that characterize most organizational decision-making: time pressure, high stakes, and informational overload.

Skitka, Mosier, and Burdick (1999) demonstrated experimentally that operators using automated decision aids were more likely to make errors of both commission (acting on incorrect automated advice) and omission (failing to notice problems the automation did not flag). The automation did not simply assist the operators. It restructured their attention, reducing their vigilance for information outside the system's frame.

Logg, Minson, and Moore (2019) found in a series of experiments that people often gave greater weight to advice labeled as algorithmic than to advice labeled as human, especially in lay forecasting contexts, though this tendency was not uniform across conditions. This preference is

understandable when algorithms are well calibrated and the decision environment is stable. But in organizational contexts characterized by ambiguity and novelty, automatic deference to algorithmic outputs suppresses exactly the kind of skeptical, situated judgment the situation demands.

The deeper problem is institutional, not just individual. When an AI system produces a score or recommendation, it creates a default. Once algorithmic outputs become embedded in organizational routines, they can function as defaults that are costly to contest, even though professionals often respond with buffering strategies, critique, or selective decoupling rather than simple compliance (Christin, 2017). This dynamic does not enhance rationality. It compresses the decision space around the machine's output and penalizes independent judgment.

Consider risk assessment tools in criminal justice. Algorithmic risk scores were introduced to reduce human bias and increase consistency in pretrial and sentencing decisions. Green and Chen (2019) showed that algorithm-in-the-loop decision-making has significant limits: adding algorithmic assessments did not straightforwardly improve human performance, and explanatory support did not eliminate those limits. The human-algorithm interaction did not produce the best of both worlds. It produced a hybrid in which the algorithm's frame could dominate without the algorithm bearing responsibility for the outcome.

Decontextualization

AI systems process data. Data is, by definition, information abstracted from its original context. A medical record is not a patient. A resume is not a person. A credit history is not a life. This abstraction is necessary for computation. It is also a source of systematic error.

Selbst et al. (2019) identified five 'traps' that arise when algorithmic fairness is pursued without attention to social context. Among the most relevant is what they termed the 'formalism trap' – the assumption that a social problem can be solved by formalizing it in mathematical terms. Formalization requires simplification. Simplification requires deciding what to include and what to discard. Those decisions are themselves value-laden, yet they are typically treated as technical design choices, invisible to the decision-makers who use the system's outputs.

Alkhatib and Bernstein (2019) extended the analysis by applying Lipsky's concept of street-level bureaucracy to algorithmic systems. Traditional bureaucrats exercised discretion – sometimes well, sometimes badly, but always in response to the specifics of a case. Algorithmic systems replace this discretion with classification. The individual case becomes an instance of a

category, and the system acts on the category. When the category fits, this is efficient. When it does not – when the case is unusual, borderline, or novel – the system fails in ways that are difficult to detect from within its own logic.

Eubanks (2018) documented how automated eligibility systems in American welfare programs denied benefits to people who clearly qualified but whose situations did not fit the system's categorical structure. A person who was homeless, for example, might lack the fixed address required by the intake algorithm. The system did not recognize homelessness as a reason to adapt its categories. It simply applied them. This is not a minor operational glitch. It is a structural failure of rationality: the system could not recognize that its own categories were inadequate to the situation it faced.

Organizational decisions that matter most are precisely those that resist clean categorization. Strategic choices, ethical dilemmas, novel situations, contested priorities – these require the decision-maker to interpret the situation, to ask what kind of problem this is before reaching for a solution. AI systems, by contrast, begin with a fixed problem definition and optimize within it. They cannot step back and ask whether the problem has been correctly framed.

The suppression of reflexive judgment

The three mechanisms above converge on a single, compounding effect: AI tends to suppress the reflexive dimension of organizational judgment. By reflexive judgment, I mean the capacity to examine one's own decision premises – to ask not just 'are we achieving our target?' but 'is this the right target?' and 'do we still understand what the target means?'

This is what distinguishes rationality from mere calculation. A calculator is not rational. It performs operations on inputs it cannot evaluate. Organizational rationality, in the fuller sense developed by Simon, March, and the subsequent tradition, involves the ability to reconsider objectives, revise categories, and learn from anomalies. When AI systems are integrated into decision processes, this reflexive capacity often atrophies. The system provides answers. The answers are precise. The precision discourages questioning.

Kellogg, Valentine, and Christin (2020) reviewed extensive evidence on how algorithms restructure workplace control. They found that algorithmic management systems do not simply automate existing decisions. They redefine what counts as a decision, who gets to make it, and on what basis. This restructuring concentrates authority in the system's design – in the choice of

variables, the training data, the objective function – while dispersing responsibility across the organization in ways that make accountability difficult.

Pasquale (2015) described this as the logic of the 'black box society' – a world in which consequential decisions are made by opaque systems whose internal workings are inaccessible to the people affected by them. Opacity alone does not destroy rationality. Many institutional processes are opaque. What destroys rationality is opacity combined with the presumption of objectivity: the assumption that because the system is formal and data-driven, its outputs need not be interrogated with the same skepticism one would apply to a human judgment.

Organizational consequences

When organizations substitute optimization for judgment, several things go wrong – not dramatically, not all at once, but cumulatively and in ways that erode institutional intelligence.

Accountability weakens. Elish (2019) introduced the concept of 'moral crumple zones' to describe situations in which human operators absorb blame for failures that are structurally produced by automated systems. When an AI-assisted decision goes wrong, organizations face a characteristic dilemma: the human decision-maker technically had the authority to override the system but lacked the information, time, or institutional support to do so. Responsibility collapses onto the human, even though the decision structure was shaped by the algorithm. The result is accountability without genuine agency.

Contestability diminishes. Algorithmic decisions are difficult to challenge, partly because of technical opacity and partly because the decision logic is distributed across data, model, and implementation choices that no single person controls. An employee denied a promotion by a scoring system cannot easily argue that the scoring system is wrong, because the system's logic is not presented as an argument. It is presented as a measurement. Measurements feel factual. Arguments invite counter-arguments. This is why algorithmic authority is more resistant to challenge than human authority, even when it is less well founded.

Adaptive learning suffers. Organizations learn by noticing anomalies, questioning assumptions, and revising their models of the world. AI systems, by contrast, optimize within their given model. When an anomaly arises, the system does not learn from it in the organizational sense – it either classifies it within existing categories, which may be wrong, or flags it as noise, which may be informative. The organization that relies on its algorithmic systems for decision support becomes progressively less able to detect the situations in which those systems are misleading.

There is an irony here worth stating plainly. AI is adopted to improve organizational decisions. Its adoption often makes organizations worse at recognizing when their decisions need improvement.

Conclusion

I have argued that AI diminishes decision rationality in organizations through four interconnected mechanisms: proxy optimization that substitutes measurable targets for meaningful goals, automation bias that compresses judgment around algorithmic defaults, decontextualization that strips situations of the particularity that good decisions require, and the suppression of reflexive judgment that prevents organizations from questioning their own premises.

None of these mechanisms is accidental. They follow directly from what AI is: a technology that formalizes, computes, and optimizes. Formalization is powerful. It is also reductive. When organizations treat AI-generated outputs as rational simply because they are data-driven, consistent, and precise, they mistake the appearance of rationality for its substance.

Genuine organizational rationality is something harder and more human than optimization. It involves interpretation, contextual sensitivity, normative evaluation, and the willingness to revise one's own assumptions. These capacities are not enhanced by AI systems. In many cases, they are displaced by them.

The practical implication is not that organizations should reject AI, but that they should stop treating it as a rationality upgrade. It is a powerful tool for narrow computational tasks. It is a poor substitute for organizational judgment. The organizations most at risk are those that have forgotten the difference.

References

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: The simple economics of artificial intelligence*. Boston: Harvard Business Review Press.

Alkhatib, A., & Bernstein, M. (2019). Street-level algorithms: A theory at the gaps between policy and decisions. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Paper 530). New York: ACM. https://doi.org/10.1145/3290605.3300760

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, *2*(1), 67–90. https://doi.org/10.1016/0149-7189(79)90048-X

Christin, A. (2017). Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society*, *4*(2), Article 2053951717718855. https://doi.org/10.1177/2053951717718855

Cohen, M. D., March, J. G., & Olsen, J. P. (1972). A garbage can model of organizational choice. *Administrative Science Quarterly*, *17*(1), 1–25.

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. Cambridge, MA, & London: MIT Press.

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society*, *5*, 40–60. https://doi.org/10.17351/ests2019.260

Espeland, W. N., & Sauder, M. (2007). Rankings and reactivity: How public measures recreate social worlds. *American Journal of Sociology*, *113*(1), 1–40. https://doi.org/10.1086/517897

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York: St. Martin's Press.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107–143. https://doi.org/10.1111/j.1756-8765.2008.01006.x

Goodhart, C. A. E. (1975). Problems of monetary management: The U.K. experience. In: *Papers in Monetary Economics* (Vol. 1, pp. 1–20). Sydney: Reserve Bank of Australia.

Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), Article 50, 1–24. https://doi.org/10.1145/3359152

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

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

March, J. G. (1994). *A primer on decision making: How decisions happen*. New York: Free Press.

Muller, J. Z. (2018). *The tyranny of metrics*. Princeton, & Oxford: Princeton University Press.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. New York: Crown.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information.* Cambridge, MA, & London: Harvard University Press.

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 469–481). New York: ACM. https://doi.org/10.1145/3351095.3372828

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59–68). New York: ACM. https://doi.org/10.1145/3287560.3287598

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, *69*(1), 99–118. https://doi.org/10.2307/1884852

Simon, H. A. (1997). *Administrative behavior: A study of decision-making processes in administrative organizations* (4th ed.). New York: Free Press.

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, *51*(5), 991–1006. https://doi.org/10.1006/ijhc.1999.0252

Strathern, M. (1997). 'Improving ratings': Audit in the British University system. *European Review*, *5*(3), 305–321.