Collective and Individual Decisions
e-ISSN: 3071-7973

Rationality, Augmentation, and the Limits of Critique

Aleksandra Przegalinska[1]

Abstract

Dariusz Jemielniak's article presents a carefully constructed and valuably provocative case that AI systematically erodes organizational rationality. The argument is lucid, well-referenced, and arrives at a conclusion that many practitioners would do well to take seriously. This response, however, suggests that the paper's framing rests on a characterization of AI epistemology that the research community has substantially moved beyond, and that it measures AI's limitations against an account of human organizational judgment that is more idealized than the empirical record supports. By engaging critically with the paper's core moves, this response does not seek to minimize the real problems Jemielniak identifies, but rather to redirect the inquiry toward the more productive questions of design, governance, and institutional structure that those problems demand. The argument proceeds from a shared concern: not AI versus judgment, but the ongoing challenge of building sociotechnical systems in which human and machine capacities genuinely complement each other.

Keywords: organizational rationality, human-AI collaboration, automation bias, algorithmic decision-making, explainable AI, bounded rationality, AI governance, sociotechnical systems

[1] Aleksandra Przegalinska – Kozminski University, Warsaw, Poland, e-mail: aprzegalinska@kozminski.edu.pl, https://orcid.org/0000-0003-0864-8007

On the construction of the argument's target

It is worth beginning by reconstructing what Jemielniak's argument actually requires. His first move is to attribute to contemporary organizational discourse a "widespread faith" that AI "necessarily produces better decisions." From this premise the rest of the paper follows: if organizations believe AI is inherently rational, and if AI has properties that systematically undermine genuine rationality, then AI is making organizations worse at deciding.

The premise, however, warrants closer examination. The "widespread faith" is attributed to no specific actor, no organization, and no identified research program. It functions as a general claim about the intellectual atmosphere rather than a position held by any particular scholarly community. When one surveys the actual landscape of AI research and practice over the past decade, one finds not confident assertions of AI's inherent rationality but rather a sprawling literature in AI ethics, responsible AI, explainable artificial intelligence (XAI), algorithmic accountability, and human-AI teaming that is organized almost entirely around the proposition that AI systems require critical scrutiny, contextual embedding, and human oversight precisely because they do not automatically produce better outcomes.

The works Jemielniak himself cites, including O'Neil (2016), Eubanks (2018), Pasquale (2015), Obermeyer et al. (2019), and Selbst et al. (2019), are all products of a critical AI studies tradition whose central purpose is to challenge naive techno-optimism about algorithmic decision-making. These are not authors arguing that AI is rational. They are authors arguing, as Jemielniak does, that it is not. The paper, in other words, constructs a target and then defeats it using the very intellectual tradition that had already done so. This is worth noting not as a dismissal of the paper's contribution, but as a way of clarifying where its real novelty would need to lie.

To attribute to "contemporary organizational thinking" a faith that AI produces better decisions by definition may reflect an accurate reading of vendor communications and managerial enthusiasm, but it understates the sophistication of the scholarly conversation the paper ostensibly addresses. Acknowledging this distinction would strengthen rather than weaken the paper's core concerns.

Simon's legacy revisited

Jemielniak's reading of Herbert Simon is productive, but it draws on one dimension of Simon's legacy while setting aside another that is equally significant. Simon's bounded rationality is

invoked to argue that human cognitive limitations are not deficiencies but adaptations, that satisficing heuristics are often well-calibrated to their environments, and that good organizational judgment requires capacities that optimization alone cannot replicate. This is a defensible reading.

It is, however, a partial one. Simon spent the latter decades of his career at the forefront of artificial intelligence research, co-authoring foundational work on information-processing theories of cognition and contributing substantially to the development of computational models of problem-solving (Newell, & Simon, 1972). His vision of bounded rationality was not intended as a defense of human intuition against machines. It was a framework for understanding cognition (biological and artificial alike) as adaptive computation under constraint. Simon was deeply interested in how computational systems could extend and augment human rational capacities, rather than simply replicate or displace them.

The garbage can model (Cohen, March, & Olsen, 1972), cited by Jemielniak as evidence of the irreducible complexity of organizational decision-making, deserves a similarly nuanced reading. The model describes organizations characterized by ambiguous goals, unclear causal structures, and decision processes that couple problems and solutions in ways that are difficult to anticipate or control. Jemielniak reads this as evidence that organizations require interpretive flexibility and contextual sensitivity, which is quite right. But it also describes organizations that are capable of making poor decisions for politically motivated, structurally distorted, and cognitively biased reasons. A balanced account of what human organizational judgment actually delivers, drawing on the extensive literature documenting institutional dysfunction, conformity pressures, motivated reasoning, and availability bias (Kahneman, 2011; Pfeffer, 1981), would provide a more honest baseline against which AI's performance can be assessed.

Reflexive judgment: An asymmetric account

The paper's most philosophically ambitious claim is that AI suppresses reflexive judgment, understood as the capacity to question one's own premises, revise objectives, and recognize when established categories no longer fit the situation. This is a real and important phenomenon in poorly designed AI deployments, and it deserves the attention Jemielniak gives it. It is worth noting, however, that the account is asymmetric: the same suppression of reflexivity is well documented in human organizations, and has been for decades.

Argyris and Schön’s (1978) foundational work on organizational learning documented the pervasiveness of what they called single-loop learning, the tendency of organizations to

respond to errors by adjusting behavior within existing norms rather than questioning the norms themselves. Weick (1995) analyzed how organizational sensemaking produces collective cognitive maps that filter out anomalies, normalize deviance, and make catastrophic failure possible precisely because human judgment operates within interpretive frames that feel self-evidently correct. Vaughan's (1996) study of the Challenger disaster remains among the most detailed accounts produced of how human organizational processes, embedded in institutional cultures and bureaucratic routines, can systematically suppress the reflexive recognition that something is going wrong.

Jemielniak argues that AI systems cannot step back and ask whether the problem has been correctly framed. This is true of current narrow AI systems in specific deployment contexts. The same difficulty, however, characterizes many human decision-makers in many organizational settings. The more productive question is not whether AI has this limitation (it does) but whether the overall sociotechnical system, combining human and algorithmic capacities, is better or worse at reflexive learning than fully human systems. That is an empirical question, and the evidence is genuinely mixed rather than uniformly negative.

Indeed, there are documented cases in which the introduction of algorithmic systems has enhanced organizational reflexivity by making decision patterns visible, auditable, and comparable in ways that purely discretionary human judgment resists. Wachter, Mittelstadt, and Russell (2017) argue that formal algorithmic systems create new surfaces for contestation and accountability precisely because their logic can be examined and challenged. This does not make algorithmic systems unproblematically superior; it does suggest that the reflexivity calculus is more complicated than the paper's framing allows.

From failure cases to design principles: A methodological consideration

Jemielniak's empirical strategy draws heavily on documented failure cases: the racial bias in healthcare resource allocation (Obermeyer et al., 2019), the limits of algorithmic risk assessment in criminal justice (Green, & Chen, 2019), and the exclusions produced by automated welfare eligibility systems (Eubanks, 2018). These are important cases. They represent genuine harms and genuine failures of design, governance, and institutional oversight, and they deserve the prominent place they have in this literature.

A methodological concern arises, however, when failure cases are used to establish a general thesis about AI and organizational rationality, rather than to characterize a class of failures in

specific deployment contexts. Any technology produces failure cases. Medical evidence review processes, legal precedent systems, financial models, and human expert panels all produce documented, systematic failures traceable to structural features of how they are designed and used. The existence of such failures establishes design requirements and governance imperatives; it does not in itself support a general claim about the technology's relationship to rationality.

The healthcare algorithm analyzed by Obermeyer et al. (2019) failed because it used cost as a proxy for need in a context where race-correlated differences in healthcare utilization made that proxy systematically misleading. This is a profound and important finding. It is also, considered carefully, an argument for better proxy selection, stronger bias auditing, and more careful problem formulation, rather than an argument against algorithmic support for resource allocation as such. The move from specific failure to general concern is understandable as a rhetorical strategy, but it forecloses distinctions that a constructive research agenda would need to preserve.

The criminal justice risk assessment literature is also more nuanced than the paper's treatment suggests. Dressel and Farid (2018) showed that untrained human participants performed comparably to the COMPAS recidivism algorithm, a finding that complicates any simple narrative about algorithmic inferiority relative to human judgment. The challenge in this domain is not that algorithms perform worse than humans, but that neither algorithms nor humans are adequate to the task of predicting individual future behavior, and that any prediction system deployed in high-stakes contexts requires careful attention to what is being measured, how, and with what institutional consequences.

Engagement with the human-AI collaboration literature

A significant gap in the paper’s scholarly engagement is the extensive literature on human-AI collaboration, augmented decision-making, and complementary human-machine cognition that has developed substantially over the past decade. This is not a peripheral strand of AI research. It represents a central agenda pursued across cognitive science, human-computer interaction, organizational behavior, and AI design, with the explicit goal of preserving and enhancing the very capacities Jemielniak values (contextual sensitivity, normative reflection, interpretive flexibility) while extending computational support where it is genuinely useful.

Brynjolfsson and McAfee (2014) documented the ways in which AI augments rather than replaces human capabilities across a wide range of tasks, a theme developed further in

Daugherty and Wilson's (2018) work on collaborative intelligence. Kamar (2016) analyzed the conditions under which human-AI teams outperform either humans or algorithms alone, identifying design features of effective collaboration. Lai, Liu, & Tan (2019) investigated how AI explanations can improve human decision accuracy in complex classification tasks. Amershi et al. (2019) synthesized design guidelines for human-AI interaction that directly address the automation bias and decontextualization concerns Jemielniak raises.

The participatory design tradition has produced frameworks for involving affected communities in algorithmic system design in ways that address the value awareness Jemielniak rightly identifies as essential (Sloane et al., 2022). XAI research is specifically aimed at making machine decision logic accessible to human scrutiny, enabling the kind of reflexive interrogation the paper suggests AI suppresses (Rudin, 2019; Lipton, 2018). Algorithmic impact assessment frameworks, now embedded in regulatory requirements including the EU AI Act, institutionalize mechanisms through which organizations can question their own AI-driven premises.

Engaging with this literature would, I believe, sharpen rather than dissolve the paper's concerns. The real frontier is not whether to use AI or to preserve human judgment, but how to design sociotechnical systems in which human reflexive capacities are supported rather than displaced, and in which the complementary strengths of human and algorithmic processing are deployed in configurations appropriate to specific decision contexts. This is a research question of considerable importance, and Jemielniak's identification of failure mechanisms is a valuable contribution to it.

On the Specification of 'Genuine' Rationality

The paper's central normative concept, genuine organizational rationality, is doing substantial argumentative work without being fully specified. Jemielniak distinguishes it from mere calculation and defines it through a list of capacities: contextual sensitivity, interpretive flexibility, value awareness, and practical expertise in Dreyfus's sense. These are real and important capacities. The question of what precisely makes their exercise "genuine" in a way that algorithmic processes cannot approximate, however, is left underspecified.

The paper implies that there is a form of rationality that organizations possess when they rely on human judgment and forfeit when they adopt AI. Human organizational judgment does not, however, automatically possess the capacities attributed to it here. Organizational psychology has extensively documented how institutional pressures, professional socialization,

hierarchical authority, and political dynamics can distort exactly the capacities the paper valorizes. Contextual sensitivity can be suppressed by bureaucratic categorization. Interpretive flexibility can be constrained by standard operating procedures. Value awareness can be crowded out by performance metrics that are human-designed and human-enforced. Practical expertise is distributed unequally and frequently overridden in formal decision processes.

The alternative to AI optimization is not unmediated human wisdom. It is human judgment embedded in organizational systems that are themselves formal, constrained, politically structured, and subject to well-documented distortions. Measuring AI against an idealized account of organizational rationality rather than against the reality of organizational decision-making sets a standard that human processes themselves rarely meet. A more symmetrical comparison would strengthen the paper's most important empirical claims.

Jemielniak's practical counsel, that organizations should stop treating AI as a rationality upgrade and recognize it as a powerful tool for narrow computational tasks rather than a substitute for organizational judgment, is sensible and worth heeding. It would be more persuasive if paired with an equally candid account of the conditions under which purely human organizational judgment also falls short, and of what governance structures might improve both.

Toward a more constructive research agenda

None of the foregoing is intended to minimize the concerns Jemielniak raises. Proxy optimization is a real and serious problem. Automation bias is well documented. Decontextualization produces failures with significant human costs. The suppression of reflexive judgment in organizations that deploy AI tools is an important phenomenon that deserves sustained empirical attention and institutional resistance. These are genuine contributions, and the paper makes them clearly.

The concern here is with the level of the argument. By framing the issue as AI versus rationality rather than as a design, governance, and institutional challenge, the paper risks deflecting attention from the more actionable questions. Which AI deployment contexts are prone to which failure modes, and why? What organizational structures and practices buffer against automation bias and preserve human reflexive capacity? How should algorithmic systems be designed to complement rather than displace contextual judgment? What regulatory and accountability frameworks are most effective at preventing proxy optimization from

hardening into institutional norm? Who should be involved in the design of AI systems that affect them, and through what mechanisms?

These questions have active research programs attached to them, developed across AI ethics, XAI, participatory design, EU AI governance, and organizational studies of human-AI teaming. A scholarly intervention that engaged with this literature, generated hypotheses testable against organizational data, and offered design-oriented recommendations calibrated to specific failure modes would, I believe, have greater practical traction than a categorical indictment.

Organizations trying to govern their AI deployments responsibly need frameworks for diagnosis and improvement. The identification of failure mechanisms is an essential first step, and Jemielniak provides it. What follows from that identification, however, is not skepticism toward AI as such but a demanding and specific set of requirements for how AI systems should be designed, evaluated, governed, and embedded within organizational structures that preserve the human capacities they most need to retain.

Conclusion

Jemielniak's paper is written with care, and its documentation of failure modes in algorithmic decision-making is a genuine contribution to the organizational literature. The mechanisms it identifies (proxy optimization, automation bias, decontextualization, and the suppression of reflexive judgment) are real phenomena that deserve sustained attention. The failure cases analyzed are important and insufficiently understood by many practitioners. This is useful work.

The response offered here does not challenge the importance of these phenomena but questions the framing through which they are presented. A characterization of AI's claims to rationality that has already been substantially revised within the field, paired with an account of human organizational judgment that sets aside the extensive literature on its own failures, produces a binary that may be rhetorically effective but is analytically limiting. It forecloses the nuanced, empirically grounded, design-oriented inquiry that the problems themselves demand.

The concern, I would suggest, is not with AI per se but with the conditions of its deployment. Poorly designed, inadequately governed, and unreflectively embedded AI systems undermine organizational rationality. Well-designed systems, appropriately situated within structures that

preserve human reflexive capacity, can support it. The difference is consequential, and making it clearly is precisely what a productive research agenda in this area requires.

The organizations most at risk are those that have deployed AI without asking whether it is working as intended, for whom, and at what cost, and also those that have dismissed it on the basis of categorical skepticism rather than careful evaluation. Genuine organizational rationality, in the richer sense that Jemielniak rightly defends, requires attending to both.

References

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (paper 3). New York: ACM. https://doi.org/10.1145/3290605.3300233

Argyris, C., & Schön, D. A. (1978). *Organizational learning: A theory of action perspective*. Massachusetts, London, & Amsterdam: Addison-Wesley.

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. New York, & London: W.W. Norton.

Cohen, M. D., March, J. G., & Olsen, J. P. (1972). A garbage can model of organizational choice. *Administrative Science Quarterly*, *17*(1), 1–25. https://doi.org/10.2307/2392088

Daugherty, P. R., & Wilson, H. J. (2018). *Human + machine: Reimagining work in the age of AI*. Boston: Harvard Business Review Press.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York: St. Martin's Press. https://doi.org/10.5204/lthj.v1i0.1386

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

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kamar, E. (2016). Directions in hybrid intelligence: Complementing AI systems with human intelligence. In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (pp. 4070–4073). New York: AAAI Press.

Lai, V., Liu, H., & Tan, C. (2019). 'Why is Chicago deceptive?' Towards building model-driven tutorials for humans. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York: ACM. https://doi.org/10.1145/3313831.3376873

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, *16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs: Prentice-Hall.

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.

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

Pfeffer, J. (1981). *Power in organizations*. Marshfield: Pitman.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

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.

Sloane, M., Moss, E., Awomolo, O., & Forlano, L. (2022). Participation is not a design fix for machine learning. In: *Proceedings of the 2022 ACM FAccT Conference* (pp. 1161–1174). New York: ACM. https://doi.org/10.48550/arXiv.2007.02423

Vaughan, D. (1996). *The Challenger launch decision: Risky technology, culture, and deviance at NASA*. Chicago, & London: University of Chicago Press.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, *31*(2), 841–887. https://doi.org/10.2139/ssrn.3063289

Weick, K. E. (1995). *Sensemaking in organizations*. Thousand Oaks: Sage. https://doi.org/10.1016/S0956-5221(97)86666-3