Algorithmic Conformity and the Limits of Human-in-the-Loop

By Yotam Liel (Cornell Tech)

Algorithmic Conformity and the Limits of Human-in-the-Loop

By Yotam Liel (Cornell Tech)

Artificial intelligence (AI) now influences decisions with critical consequences for people's lives: who gets hired, who receives bail, which patients get which treatments. These systems can improve decision-making in important ways, but they may also exhibit systematic biases, make errors, and fail in unexpected ways with serious consequences.

One dominant safeguard against these failures is keeping a human in the loop. Let AI assist, but ensure a human holds the final decision. We expect this human to exercise independent judgment, catch the algorithm's errors, override problematic recommendations. But what if this assumption is wrong?

In a recently published paper (1), we ask: To what extent can we trust humans in the loop to remain appropriately critical of AI advice? We argue that this isn't just a question about human competence compared to that of the AI, it's also a question about the types of pressures created by those sociotechnical systems.

Indeed, we've already seen troubling evidence of people following erroneous AI advice across domains. For example, in one widely reported incident, a lawyer relied on ChatGPT for legal research and presented fabricated court cases to a judge (2). In another, highly troubling case, healthcare workers report situations where AI-generated alerts override their clinical judgment, pushing them to follow protocols they believe may not be in patients' best interests (3).

Existing attempts to explain AI overreliance have primarily focused on AI's informational influence, where people defer to AI because they believe it has superior analytical capabilities or access to better information. This line of rationalization, albeit problematic, could make sense in contexts of high complexity or uncertainty, where human judgment genuinely stands to benefit from AI assistance.

Our research extends current explanations of overreliance on AI by proposing and testing an additional mechanism: normative pressure. We argue that people conform to AI recommendations not just because they think the AI is more competent (informational influence), but also because the organizational and social embedding of AI creates perceived expectations and obligations to defer to its guidance. In other words, people sometimes follow AI advice because they feel they should, even when they privately disagree with it.

Our theory builds on the Computers Are Social Actors (CASA) paradigm, and on the concept of algorithmic authority. CASA studies have shown that people automatically apply social rules and norms to technology, even while consciously knowing these systems are not human (4). Meanwhile, research on algorithmic authority shows that algorithms gain authority not merely through computational accuracy, but also through their embedding within social and organizational processes (5–7).

This institutional embedding generates normative influence through multiple channels. First, organizational deployment of AI signals institutional endorsement, leading workers to see these systems as representing institutional authority. Second, professional norms equate technological adoption with competence, such that adopting AI becomes part of being seen as a modern professional. Third, formal policies integrate AI into official workflows, creating potential consequences for those who deviate from algorithmic recommendations.

To test whether people experience normative pressure from AI, we designed experiments inspired by Solomon Asch's classic social conformity studies (8). A key feature of those studies was that participants performed tasks where the correct answer is obvious, yet they faced contradictory information from peers. This design isolates social pressure from genuine uncertainty about the right answer.

We adapted this approach for AI: participants performed straightforward image classification tasks where they could easily determine the correct answer independently. These were intentionally simple tasks that any person can perform accurately without assistance. The critical manipulation was that participants received AI recommendations that were clearly wrong. We conducted four studies with 1,445 crowdworkers simulating realistic platform-work scenarios (data annotation tasks are common on these platforms).

This design allows us to distinguish between different types of AI overreliance. If people follow erroneous AI advice on complex, ambiguous tasks, that might reflect reasonable informational influence (the AI might genuinely know something they don't). But if people follow obviously wrong AI advice on simple tasks they can perform perfectly on their own, something else must be at work.

The results are strong and consistent. Between 19% and 27% of participant responses aligned with the AI's erroneous recommendations, a sharp contrast to control groups who received no advice and made these same errors less than 1% of the time. Additionally, people conformed more to AI advice than to identical advice presented as coming from other humans.

Importantly, we found evidence that normative pressure operates independently of informational influence. We measured both participants' confidence in the AI's accuracy and their sense of pressure or discomfort when considering disagreeing with the AI. While we found that both factors predicted conformity, normative pressure drove people to follow

AI advice even among participants who reported low confidence in the AI's recommendations.

Encouragingly, we also identified a way to reduce, although not completely mitigate, conformity to AI. When participants believed their decisions would impact autonomous vehicle safety (high stakes), conformity dropped by half as compared to when the same tasks were framed as affecting warehouse efficiency (lower stakes). This suggests that emphasizing the moral weight and real-world consequences of decisions can help people resist normative pressure to defer to AI.

Our studies also rule out alternative explanations. In one study, we enforced minimum task completion times to ensure people weren't rushing through without thinking. And in another, we offered financial incentives for accuracy. Neither mitigated conformity. Additionally, our results show that those who conformed to the AI's erroneous advice spent significantly more time on the tasks than those in a control group who received no advice. These findings exclude explanations such as inattention, careless behavior, and accepting the AI advice to rush through the tasks.

Our findings reveal a critical challenge for how we govern AI systems. Current approaches assume that keeping humans in the loop provides a safeguard against algorithmic errors and biases. But our research shows that organizational and social structures can create pressures that undermine this safeguard.

This connects directly to what Cooper and colleagues (9), updating Helen Nissenbaum's work on accountability (10), identify as "the human user as scapegoat." They describe how organizations, such as manufacturers and owners of AI-enabled systems, evade liability for harms by shifting blame to humans-in-the-loop. Our findings further emphasize the vulnerability of the human user in this process: organizations embed AI in ways that create normative pressure on humans to defer, then blame them when failures occur.

An AI tool in an organizational context is never just providing information. It also carries normative weight, creating perceived expectations about deference. This has concrete implications. Organizations must examine how their policies, workflows, and professional cultures position AI systems. Do they signal that human judgment is not just permitted but expected, or do they pressure users to defer?

References

1. Y. Liel, L. Zalmanson, Turning Off Your Better Judgment: Algorithmic Conformity in Artificial Intelligence-Human Collaboration. *Journal of Management Information Systems* 42, 1087–1117 (2025).

2. B. Weiser, ChatGPT Lawyers Are Ordered to Consider Seeking Forgiveness. *The New York Times* (2023).

3. S. Banker, S. Khetani, Algorithm Overdependence: How the Use of Algorithmic Recommendation Systems Can Increase Risks to Consumer Well-Being. *Journal of Public Policy & Marketing* 38, 500–515 (2019).

4. C. Nass, Y. Moon, Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56, 81–103 (2000).

5. C. Shirky, A Speculative Post on the Idea of Algorithmic Authority. (2009). Available at: https://web.archive.org/web/20190224040538/http://www.shirky.com/weblog/2009/11/a-speculative-post-on-the-idea-of-algorithmic-authority/ [Accessed 5 December 2024].

6. T. Gillespie, “The Relevance of Algorithms” in *Media Technologies*, T. Gillespie, P. J. Boczkowski, K. A. Foot, Eds. (The MIT Press, 2014), pp. 167–194.

7. C. Lustig, B. Nardi, Algorithmic Authority: The Case of Bitcoin in *2015 48th Hawaii International Conference on System Sciences*, (IEEE, 2015), pp. 743–752.

8. S. E. Asch, Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied* 70, 1–70 (1956).

9. A. F. Cooper, E. Moss, B. Laufer, H. Nissenbaum, Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning in *2022 ACM Conference on Fairness Accountability and Transparency*, (ACM, 2022), pp. 864–876.

10. H. Nissenbaum, Accountability in a computerized society. *Sci Eng Ethics* 2, 25–42 (1996).

Yotam Liel
Postdoctoral Fellow
Cornell Tech
yl4363@cornell.edu
More Info >