Journal of the Statistical and Social Inquiry Society of Ireland
Vol. LIV

Symposium: Big Tech and Big Data - New Frontiers in Regulation

The Human in the Machine: Power, Bias, and Governance in AI Societies

Taha Yasseri[1]

Trinity College Dublin
Technological University Dublin
University College Dublin

(read before the Society, 30th April 2025)

Abstract: As artificial intelligence (AI) systems become embedded in everyday life, they increasingly participate in decisions, interactions, and institutional processes once governed solely by humans. This article examines the evolving role of AI not as a neutral tool, but as a socio-technical agent shaped by and shaping human norms, values, and structures of power. Drawing on insights from computational sociology, behavioural experiments, and human-machine collaboration, we explore how gender bias, trust asymmetries, and algorithmic governance unfold across domains - from digital assistants and workplace management to collective intelligence and online platforms. Through case studies, including large-scale experiments and Wikipedia-based modelling, we illustrate the dynamics of cooperation, conflict, and consensus in hybrid human-machine systems. We argue that ethical design and regulation must move beyond principles to address structural inclusion, institutional accountability, and sociotechnical transparency. By situating AI within broader social and political contexts, we offer a framework for understanding and shaping its impact on human autonomy, fairness, and collaboration. The future of AI, we contend, is not determined by technical capacity alone, but by the values and institutions that govern its development and deployment.

Keywords: Artificial Intelligence, Algorithmic Bias, Human-Machine Interaction, Gender, Collective Intelligence
JELs: O33, D83, J71, L86

1. INTRODUCTION

In recent years, the pervasive integration of artificial intelligence (AI) into everyday life has sparked a re-evaluation of long-standing assumptions about human agency, autonomy, and social structure. Far from being confined to laboratories or speculative fiction, AI now influences how we work, communicate, learn, govern, and relate to one another. From algorithmic hiring tools and predictive policing to conversational agents and content moderators, intelligent systems are no longer peripheral to society; they are constitutive of it. This shift from auxiliary automation to core participation demands a rethinking of the human-machine relationship not merely as a technical configuration but as a fundamentally social phenomenon (Crawford 2021).

AI systems today are designed not only to execute tasks but also to learn, adapt, and make autonomous decisions based on probabilistic reasoning and massive datasets. While this promises efficiency and scalability, it also raises profound questions about fairness, accountability, and control. The very characteristics that make machine learning effective - opacity, scalability, and responsiveness to data - can simultaneously undermine democratic values and human dignity when left unexamined (Pasquale 2015; Jobin, Ienca, and Vayena 2019). The notion that

[1] ORCID: 0000-0002-1800-6094 - Email: taha.yasseri@tcd.ie. The research conducted in this publication was funded by the Irish Research Council under grant number IRCLA/2022/3217, ANNETTE (Artificial Intelligence Enhanced Collective Intelligence). We thank Workday for financial support.

algorithms are "neutral" or "objective" has been widely debunked; these systems are embedded with the assumptions, biases, and power relations of their developers, training data, and deployment contexts (Barocas, Hardt, and Narayanan 2019).

At the heart of these debates lies the question: what does it mean to be human in a machine-mediated society? This is not a rhetorical flourish but a pressing inquiry as the boundaries between biological, social, and computational agents blur. AI technologies increasingly shape not only outcomes (e.g., who gets a loan or job) but also processes of perception, interaction, and meaning-making. In fields as diverse as healthcare, criminal justice, and education, algorithms participate in shaping norms, values, and social roles, often invisibly and without recourse for those affected (Floridi and Cowls 2019; Nissenbaum 2010).

Moreover, AI does not enter society as a *tabula rasa*. It is developed, implemented, and interpreted within existing structures of inequality and ideology. As such, it can both challenge and reinforce systems of patriarchy, racism, classism, and ableism. The gendering of voice assistants, the racialized misidentification in facial recognition, and the reinforcement of stereotypes in language models exemplify the reproduction of social hierarchies through technology (Crawford 2021). These are not incidental bugs; they are systemic features reflecting the data and design choices underlying these technologies.

To understand AI's impact, we must move beyond isolated critiques of bias or error and engage with the broader socio-technical assemblages in which these systems operate. This requires an interdisciplinary approach that combines insights from computer science, sociology, philosophy, and law. It also demands a methodological pluralism that includes computational modelling, ethnography, behavioural experiments, and institutional analysis (Rahwan 2018; Tufekci 2015).

This paper discusses a computational sociology of human-machine systems (Yasseri, 2024), one that attends to power, culture, and interaction while remaining attentive to the affordances and constraints of emerging technologies. We seek not to vilify or valorize AI, but to illuminate its role in mediating human relations and institutional arrangements. In doing so, we aim to foreground the importance of governance, design, and public deliberation in shaping the future of AI in society.

2. FROM TOOLS TO AGENTS: THE EVOLUTION OF AI IN SOCIETY

2.1 Historical Context: From Automation to Autonomous Agents

The history of artificial intelligence is often framed as a series of technical milestones: from early symbolic reasoning and rule-based systems to the advent of machine learning and neural networks. Yet, equally important is the evolving *social role* of machines. In the industrial era, machines were largely conceptualized as extensions of the human body: tools that amplified strength and efficiency but remained subordinate to human control (Noble 1984). The automation of mechanical labour redefined productivity and economic structure, but the underlying premise remained one of tool use: the human operator was central, and the machine was instrumental.

This paradigm began to shift with the emergence of computing. The programmable computer introduced a new category of machine, one capable of conditional logic, memory, and abstraction. The automation of *cognitive* tasks, not just physical ones, ushered in what Wiener (1950) described as a "cybernetic revolution," where feedback, adaptation, and decision-making could be embedded within technical systems. As algorithmic systems grew in complexity, the line between user and tool became increasingly ambiguous.

The contemporary era of machine learning and large-scale models has accelerated this transformation. AI systems now not only process information but also detect patterns, infer preferences, and make predictions in real time. From self-driving vehicles to conversational agents and recommendation engines, AI technologies are no longer merely tools, they operate as *agents* in sociotechnical systems. These systems act with a form of goal-oriented autonomy, albeit bounded by data and parameters set by their developers (Russell and Norvig 2021).

This *agentic* quality of AI has significant implications. It challenges classical distinctions between human and machine, user and artefact, subject and object. It also reconfigures notions of responsibility: when outcomes are determined by systems that evolve over time through data exposure, the question of "who is accountable" becomes murky (Kroll et al. 2017). The machine is no longer a passive conduit; it is an active participant in shaping outcomes, sometimes in ways unintended even by its designers.

2.2 AI in Decision-Making: Roles, Risks, and Social Perception

As AI systems become embedded in decision-making processes across domains - credit scoring, predictive policing, medical diagnostics, and hiring - they increasingly act as *institutional agents*. Their outputs inform,

constrain, and sometimes override human judgment. In such contexts, AI is often perceived not just as a tool, but as a source of authority (Gillespie 2018; Binns 2018). This shift has prompted growing concern about the legitimacy and trustworthiness of algorithmic decision-making.

A central risk lies in the opacity of many AI systems. Black-box models, particularly those using deep learning architectures, offer high predictive accuracy but little interpretability. When individuals are denied benefits or subjected to sanctions based on algorithmic decisions, the lack of transparency undermines procedural fairness and limits avenues for redress (Burrell 2016; Selbst and Barocas 2018). This is especially problematic in high-stakes settings such as criminal justice or healthcare, where the right to explanation and appeal is foundational.
Public perception of AI decisions is also shaped by context, framing, and anthropomorphic cues. Studies show that people ascribe greater trust to systems that appear "intelligent" or "neutral," but this trust erodes rapidly when errors are made, especially when the system lacks explainability (Shin 2021). Moreover, research reveals that people judge the same decision differently depending on whether it comes from a human or a machine, and whether the machine is anthropomorphized or not (Cui and Yasseri 2025; Lee et al. 2015).

Importantly, the perceived legitimacy of AI decisions does not depend solely on their *accuracy* but also on their *interpretability* and the *social narratives* surrounding them. An opaque but highly accurate model may be distrusted if it is seen as unaccountable or misaligned with public values (Rahwan 2018). Conversely, less accurate but more interpretable systems may enjoy higher levels of acceptance, particularly in contexts requiring moral or discretionary judgment.

In sum, the evolution of AI from tool to agent marks a profound shift in the socio-technical landscape. It demands not only new technical safeguards but also a re-examination of our normative frameworks for responsibility, legitimacy, and trust in decision-making processes mediated by machines.

3. THE GENDERED MACHINE

3.1 Anthropomorphism and Default Gendering of AI

As artificial intelligence systems become more socially integrated, they are increasingly designed - and perceived - as social actors. This design tendency is underpinned by anthropomorphism: the attribution of human traits, emotions, or intentions to non-human entities. While anthropomorphism can enhance user engagement and system usability, it also introduces new social and psychological dynamics, particularly around gender (Nass and Moon 2000; Reeves and Nass 1996).

Digital assistants like Siri, Alexa, and Cortana have been launched with default female voices and names, reinforcing cultural stereotypes of women as helpful, compliant, and non-threatening. Scholars have argued that this reflects and perpetuates a "submissive femininity" in service technologies (West et al. 2019). These design choices are rarely neutral; they embed social norms and role expectations into the architecture of AI systems. Even when gender is not explicitly specified, users frequently project gender onto AI agents based on voice, language, behaviour, or perceived functionality (Wong and Kim 2023).

This default gendering has far-reaching implications. It not only shapes user expectations and interaction patterns but also amplifies existing social biases in algorithmic mediation. For instance, users tend to demand more emotional labour from female-voiced AI systems and are more forgiving of errors made by male-voiced counterparts (Mahmood and Huang 2023). Such asymmetries replicate gendered patterns of responsibility and emotional burden that exist in human relationships.

3.2 Gender Bias in AI Interaction

Gender norms do not merely inform how AI systems are designed - they also mediate how they are perceived and interacted with. A growing body of empirical research has shown that users' responses to AI agents vary significantly depending on the agents' perceived or assigned gender. These differences are particularly pronounced in contexts of cooperation, trust, and punishment (Gambino et al. 2020; Shen et al. 2021).

When AI systems are gendered female, users tend to associate them with warmth, cooperation, and emotional responsiveness. Conversely, male-coded systems are perceived as more competent and authoritative (Nass, Moon, and Green 2006). These gendered expectations shape user behaviour. For example, Bazazi, Karpus, and Yasseri (2024) found that participants were significantly more likely to exploit female-labelled AI agents in cooperative games, even when those agents behaved helpfully, suggesting an opportunistic bias. Male-labelled agents, by contrast, elicited defensive strategies and greater distrust in competitive contexts. These patterns mirror those seen in human-human interactions and highlight the transfer of social expectations to non-human entities.

The persistence of such behaviours reflects broader gender schemas rooted in societal norms. Women are often expected to be cooperative and self-sacrificing, while men are stereotyped as assertive and rational (Eagly and Karau 2002). The fact that users impose these same roles on AI agents - regardless of the agents' actual behaviour - illustrates how cognitive shortcuts and cultural stereotypes operate even in technologically novel environments. It also underscores a deeper concern: the replication and reinforcement of social hierarchies in human-AI systems.

These biases extend beyond cooperation into perceptions of fairness and authority. In experiments involving AI managers, participants judged decisions made by female-labelled agents as less fair, especially in punitive contexts (Cui and Yasseri 2025). This reaction reflects a normative double bind - when female agents act assertively or enforce rules, they violate gendered expectations and are penalized for doing so. This phenomenon closely parallels backlash effects documented in human leadership, where women in authority roles are evaluated more negatively for the same behaviours that benefit men (Rudman and Glick 2001).

Together, these findings point to a clear conclusion: the introduction of gender into AI design does not neutralize social biases, it reactivates and extends them. As such, designers and policymakers must confront the socio-political consequences of anthropomorphic design, especially when it intersects with gendered stereotypes and authority roles.

4. AI AS COLLEAGUE AND MANAGER

4.1 Algorithmic Governance in the Workplace

The workplace has emerged as a primary arena for the deployment of AI systems - not merely to support workers, but to govern them. From logistics and manufacturing to finance, healthcare, and the gig economy, algorithmic management has become a structural feature of organizational life (Mateescu and Nguyen 2019). These systems allocate tasks, monitor performance, assign bonuses or penalties, and even make hiring and firing decisions, often with little human oversight.

This transformation marks a shift from traditional managerial authority to *algorithmic governance*, wherein computational systems codify and enforce rules that were once socially negotiated (Lee et al. 2015; Kellogg, Valentine, and Christin 2020). The appeal of such systems lies in their purported neutrality and efficiency. Employers cite data-driven decision-making as a way to reduce bias and human error. Yet, as scholars have shown, algorithmic governance does not eliminate bias - it often operationalizes it in more opaque and scalable ways (Barocas, Hardt, and Narayanan 2019; Binns 2020).

Crucially, workers subject to algorithmic control frequently report feeling dehumanized and surveilled. In platforms like Uber, Deliveroo, and Amazon warehouses, performance metrics are tracked continuously and fed into algorithmic models that determine work schedules, wages, and job security (Rosenblat and Stark 2016; Gray and Suri 2019). These systems often leave little room for negotiation or contestation, creating environments of asymmetrical power where the algorithm is both omniscient and unaccountable.

A central concern in algorithmic management is the question of fairness. Not only in outcomes but in processes. Workers increasingly find themselves subject to decisions they cannot understand, question, or appeal. This lack of *explainability* undermines perceived legitimacy, particularly when AI systems produce adverse outcomes such as denied promotions or disciplinary actions (Burrell 2016; Shin 2021). Explainability is not just a technical feature, it is a *social requirement* for legitimacy.

4.2 Delegated Accountability and Power Asymmetries

The deployment of AI in managerial roles also complicates accountability. When an AI system makes a questionable or harmful decision, who is responsible? The software vendor? The data scientist? The employer who implemented the system? This diffusion of responsibility creates a responsibility vacuum, in which no actor is clearly liable (Pasquale, 2015).

This is especially problematic when algorithmic decisions disproportionately affect marginalized groups. Without clear lines of accountability, redress becomes difficult, and structural inequalities are reproduced through technological proxies. Even when companies include “humans in the loop,” these individuals may have little actual authority to override or interrogate the system’s outputs (see Rahwan 2018).

Moreover, algorithmic systems tend to encode managerial logics that prioritize efficiency, predictability, and quantification. These logics often conflict with workers’ needs for autonomy, dignity, and procedural fairness. As a result, the workplace becomes a site of contested rationalities - where human values are negotiated, ignored, or overridden by computational imperatives (Zuboff 2019).

The concentration of power in algorithmic systems thus reflects and intensifies broader trends in labour precarity and corporate opacity. If left unchecked, AI will not merely manage workers, but it will redefine the very meaning of work, responsibility, and organizational life.

5. HUMAN-AI COLLECTIVE INTELLIGENCE

5.1 Enhancing (or Disrupting) Group Decision-Making

The concept of *collective intelligence* - the enhanced capacity that emerges from group collaboration - has long been studied in social science, organizational behaviour, and cognitive psychology (Woolley et al. 2010). With the rise of AI, particularly large language models (LLMs), the notion has evolved from human-only collectives to *hybrid systems* where humans and machines jointly generate, validate, and act upon knowledge. These human-AI collectives hold the promise of improving group decisions by pooling cognitive resources, mitigating bias, and offering novel insights (Vaccaro et al. 2024, Cui and Yasseri 2024).

In theory, AI can support group decision-making by aggregating information, simulating outcomes, and reducing noise. For instance, AI can help surface relevant arguments, identify inconsistencies, or moderate emotional escalation in online discussions. In practice, however, the integration of AI into collective settings is far from straightforward. As recent experiments show, AI contributions are not always additive, they can dominate, distract, or distort human inputs depending on how they are introduced (Jakesch et al. 2023, Burton et al., 2024).

The effects of AI on group performance are context dependent. When AI systems are positioned as assistants, they often augment human judgment. But when presented as experts or decision-makers, they can induce overreliance or cognitive offloading, especially among less confident participants. Moreover, the presence of AI may suppress dissent, as group members defer to algorithmic "objectivity" rather than challenge flawed reasoning or implicit bias (Shin 2021; Rahwan et al. 2019).

5.2 Epistemic Authority, Alignment, and Trust

The success of human-AI collaboration hinges on *epistemic alignment* - the degree to which human users perceive the AI's contributions as credible, relevant, and intelligible. This is not simply a matter of accuracy, but of *trust*. Research shows that trust in AI systems depends on a complex interplay of performance, transparency, anthropomorphic cues, and institutional framing (Cui and Yasseri 2024).

Epistemic authority is socially constructed. Users assign credibility to AI systems based not only on prior performance but also on presentation, context, and social signalling. For example, an AI system embedded in a university platform may be granted more trust than the same system in a commercial app - even if their outputs are identical (Hannig et al. 2018). Similarly, LLMs with human-like fluency are often perceived as more knowledgeable than they are, leading to unwarranted confidence in their outputs (Floridi and Chiriatti 2020).

This misalignment can have serious consequences in domains such as journalism, education, and scientific research. If AI outputs are trusted without scrutiny, errors can propagate through human collectives. Conversely, if AI is consistently distrusted, its potential contributions are ignored or underutilized. Striking the right balance requires not just technical safeguards but also socio-technical design that fosters appropriate calibration of trust (Lee et al. 2015).

5.3 AI Agents in Collaborative Environments

The integration of AI agents into collaborative platforms such as Wikipedia, Slack, and GitHub marks a new frontier in collective intelligence. These agents can edit content, suggest revisions, flag inconsistencies, and even mediate disputes. On Wikipedia, bots already perform a large share of maintenance tasks, but as bots become more "intelligent," they increasingly participate in editorial decisions that shape public knowledge (Sumi et al. 2011; Geiger and Halfaker 2013; Tsvetkova et al. 2017).

Recent advances in LLMs raise the possibility of more autonomous AI collaborators. Experiments with "LLM societies," where thousands of AI agents simulate social media interactions, suggest that these systems can develop emergent behaviours, norms, and structures akin to human communities (Eide et al. 2016; Tsvetkova et al. 2017; Tsvetkova et al. 2024). While this opens exciting possibilities for scalable knowledge production, it also raises concerns about coherence, bias, and control.

The key challenge in these environments is *coordination*. How can human and AI agents share goals, respect boundaries, and resolve conflicts? Existing models of human collaboration - such as common knowledge and mutual intentionality - may need to be rethought or extended to accommodate machine agents with evolving behaviour (Vaccaro et al. 2024).

Ultimately, the promise of AI-enhanced collective intelligence lies not in replacing humans, but in designing systems where human judgment, ethical reasoning, and social accountability remain central. The future of collaboration is not machine-dominated, but machine-mediated - and that mediation must be intentional, inclusive, and adaptive (Mohammadi and Yasseri 2025).

6. ETHICAL AND REGULATORY PATHWAYS

6.1 Sociotechnical Design: Fairness, Transparency, Participation

The ethical challenges posed by AI are not merely abstract or philosophical - they are deeply embedded in design practices, data infrastructures, and institutional contexts. Accordingly, ethics in AI must be grounded in a *sociotechnical perspective* that acknowledges how social values are encoded into technical systems, intentionally or otherwise (Nissenbaum 2010; Barocas, Hardt, and Narayanan 2019).

This perspective reframes ethical AI not as a checklist of principles, but as an ongoing process of negotiation between competing goals, actors, and values. Design choices, such as how datasets are curated, which variables are weighted, or what outputs are emphasized, are never neutral. They reflect normative assumptions about what is important, who is valued, and what constitutes harm (Crawford 2021; Binns 2020).

Fairness, often seen as the cornerstone of ethical AI, is itself a contested concept. It can mean treating everyone the same (equality), adjusting for structural disadvantage (equity), or ensuring group-level parity (statistical fairness). Each interpretation implies different design imperatives and trade-offs (Binns 2018; Selbst et al. 2019). Therefore, fairness must be contextualized within the specific domain, stakeholder needs, and historical legacies of exclusion.

Transparency, too, is multifaceted. Technical explainability - how an AI system arrives at a decision - is only part of the puzzle. Social transparency involves disclosing who designed the system, what data were used, what values were prioritized, and how redress mechanisms are structured (Burrell 2016; Jobin, Ienca, and Vayena 2019). Without this broader transparency, AI systems risk becoming what Pasquale (2015) calls “black boxes of authority.”

Participation, finally, is crucial to ensuring that ethical design is not top-down. Involving diverse stakeholders, including marginalized communities, end-users, and civil society, enhances the legitimacy, robustness, and accountability of AI systems. Participatory design practices also surface unanticipated risks, making them more responsive and adaptive to real-world complexities (Costanza-Chock 2020; Whittaker et al. 2018).

6.2 Policy Recommendations and Governance Frameworks

Effective governance of AI requires moving beyond voluntary ethics statements and toward binding frameworks that enforce accountability. While many organizations have issued ethical guidelines, implementation has been uneven, and enforcement mechanisms are weak or absent (Jobin, Ienca, and Vayena 2019).

At the regulatory level, several jurisdictions are developing or implementing AI-specific legal instruments. The European Union’s proposed AI Act represents one of the most comprehensive efforts, classifying systems by risk level and imposing stricter requirements on high-risk applications (Veale and Borgesius 2021). Yet, even such forward-thinking regulation faces challenges in keeping pace with rapid technological development and ensuring cross-sector coherence. To complement regulation, other governance tools are gaining traction: algorithmic audits, impact assessments, transparency registers, and third-party oversight bodies. These mechanisms shift the burden of proof from individuals to institutions, requiring organizations to demonstrate that their systems are lawful, ethical, and socially beneficial (Kroll et al. 2017).

Public-sector procurement rules can also be leveraged to incentivize ethical AI. Governments are major customers of AI services - particularly in areas like healthcare, education, and law enforcement. By setting standards for transparency, fairness, and data protection in their contracts, public bodies can shape the broader market and establish benchmarks for ethical practice. Ultimately, governance must be pluralistic, combining legal, institutional, technical, and participatory approaches. It should also be iterative, allowing systems to evolve in response to new risks, contexts, and feedback loops (Yasseri, 2025).

6.3 Moving Beyond Ethics: Institutions, Power, and Inclusion

While ethics is necessary, it is not sufficient. Ethics without enforcement is performative. Ethics without power analysis is naïve. To meaningfully address the societal impact of AI, we must foreground *institutional design*, *power asymmetries*, and *political economy* (Zuboff 2019; Crawford 2021). Institutions shape how values are encoded into policy, how harms are adjudicated, and how justice is operationalized. Yet, many of the organizations

developing and deploying AI systems are private corporations with little incentive for democratic accountability. Their priorities, including profit, market dominance, and brand management, often diverge from public interest goals such as equity, inclusion, and deliberative legitimacy (Rahwan 2018; Gillespie 2018).

Moreover, power in AI is not just a matter of code or data. It is exercised through access to computational infrastructure, control over platforms, agenda-setting in policy discourse, and the capacity to shape public norms. As such, debates about ethical AI must be located within broader struggles over digital sovereignty, labour rights, surveillance, and democratic governance (Eubanks 2018; Tufekci 2015). Inclusion must be structural, not symbolic. It must address who builds AI systems, who governs them, who is protected, and who bears the risk. This requires rethinking education, labour policies, and civic infrastructure to ensure that those most affected by AI are empowered to shape its development.

In short, ethical AI is a political project. It demands not just good intentions or technical fixes, but the transformation of institutions, incentives, and imaginaries.

7. CONCLUSION

As artificial intelligence becomes deeply enmeshed in the structures of everyday life, it is imperative to resist the narrative that machines are neutral, inevitable, or autonomous in any metaphysical sense. AI is not an alien force imposed upon humanity: it is a social product, built by people, shaped by institutions, and embedded within existing power relations. Its capacities reflect the priorities, data, and assumptions of those who design and deploy it. Recognizing this social construction of AI is the first step toward reclaiming agency over its development.

Throughout this article, we have emphasized the need to keep the *human in the loop* - not as a passive overseer, but as an active shaper of socio-technical systems. This principle extends beyond individual decision points. It demands democratic engagement in the design, implementation, and governance of AI infrastructures. It requires that the values encoded in these systems be transparent, contestable, and open to revision. Importantly, "human in the loop" must not be reduced to a checkbox for compliance; it must be understood as a normative stance grounded in accountability, dignity, and participation.

The social and political effects of AI are not determined solely by technical performance, but by how systems are framed, who they serve, and how they are governed. This is why interdisciplinary approaches - drawing from computational sociology, philosophy of technology, feminist theory, and political economy - are crucial. They illuminate the ways in which AI mediates human relationships, redistributes authority, and conditions the possibilities for cooperation or conflict.

Future research must continue to explore not only *what* AI can do, but *how* it does it, *for whom*, and *at what cost*. This includes studying emergent dynamics in human-AI collectives, refining metrics for algorithmic fairness and accountability, and developing participatory models of system evaluation. Experimental platforms, agent-based simulations, and ethnographic inquiry all have a role to play in this expanded research agenda. At the same time, regulatory innovation must keep pace with technological evolution. Static rules are insufficient in the face of dynamic systems. Governance must be adaptive, inclusive, and globally coordinated - able to respond to both anticipated risks and emergent harms. Mechanisms such as algorithmic audits, impact assessments, and public algorithm registers are promising tools, but they must be supported by robust institutions, legal mandates, and civic infrastructure.

Ultimately, the question is not whether machines will surpass humans, but whether humans will retain the capacity to shape technology in ways that reflect collective values and social justice. AI can amplify human intelligence, extend democratic capabilities, and foster new forms of collaboration - but only if it is designed and governed accordingly. The path forward is neither techno-utopian nor dystopian. It is contingent, contested, and open. By cantering human values, institutional accountability, and pluralistic engagement, we can ensure that the future of AI is not only innovative but just.

References

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning*. MIT press.

Bazazi, S., Karpus, J., & Yasseri, T. (2024). AI's assigned gender affects human-AI cooperation. *arXiv preprint*, arXiv:2412.05214.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency* (pp. 149–159). Proceedings of Machine Learning Research, 81. PMLR.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 1–12.
Burton, J. W., Lopez-Lopez, E., Hechtlinger, S., Rahwan, Z., Aeschbach, S., Bakker, M. A., ... & Hertwig, R. (2024). How large language models can reshape collective intelligence. *Nature Human Behaviour*, *8*(9), 1643–1655.
Costanza-Chock, S. (2020). *Design Justice: Community-Led Practices to Build the Worlds We Need*. MIT Press.
Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.
Cui, H., & Yasseri, T. (2024). AI-enhanced collective intelligence. *Patterns*, 5(11), 100997.
Cui, H., & Yasseri, T. (2025). Gender bias in perception of human managers extends to AI managers. *arXiv preprint*, arXiv:2502.17730.
Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review, 109*(3), 573–598.
Eide, A. W., Pickering, J. B., Yasseri, T., Bravos, G., Følstad, A., Engen, V., ... & Lüders, M. (2016). Human-machine networks: Towards a typology and profiling framework. In: *Human-Computer Interaction. Theory, Design, Development and Practice* (pp. 11–22). Springer.
Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.
Floridi, L., & Chiriatti, M. (2020). *GPT-3: Its Nature, Scope, Limits, and Consequences*. *Minds and Machines*, 30(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1
Floridi, L., & Cowls, J. (2019). *A Unified Framework of Five Principles for AI in Society*. *Harvard Data Science Review*, 1(1), 2–15.
Gambino, A., Fox, J., & Ratan, R. A. (2020). *Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm*. *Human-Machine Communication*, 1, 71–86. https://doi.org/10.30658/hmc.1.5
Geiger, R. S., & Halfaker, A. (2013). When the levee breaks: Without bots, what happens to Wikipedia's quality control processes? In *Proceedings of the 9th International Symposium on Open Collaboration (OpenSym '13)* (pp. 1–6). ACM.
Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.
Gray, M. L., & Suri, S. (2019). *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt.
Hannig, L., Bush, A., Aksoy, M., Becker, S., & Ontrup, G. (2025). Campus AI vs Commercial AI: A Late-Breaking Study on How LLM As-A-Service Customizations Shape Trust and Usage Patterns. *arXiv preprint arXiv:2505.10490*.
Jakesch, M., French, M., Ma, X., & Hancock, J. T. (2023). Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). ACM.
Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.
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.
Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, 165(3), 633–705.
Lee, J. D., & See, K. A. (2004). *Trust in Automation: Designing for Appropriate Reliance*. *Human Factors*, 46(1), 50–80.
Lee, M. K., Kusbit, D., Metsky, E., & Dabbish, L. (2015). *Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers*. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1603–1612). ACM.
Mahmood, A., & Huang, C.-M. (2024). Gender biases in error mitigation by voice assistants. *Proceedings of the ACM on Human-Computer Interaction, 8*(CSCW1), 1–27.
Mateescu, A., & Nguyen, A. (2019). *Explainer: Algorithmic Management in the Workplace*. Data & Society. Retrieved from https://datasociety.net/library/explainer-algorithmic-management-in-the-workplace/
Mohammadi, S., & Yasseri, T. (2025). AI Feedback Enhances Community-Based Content Moderation through Engagement with Counterarguments. *arXiv preprint arXiv:2507.08110*
Nass, C., & Moon, Y. (2000). *Machines and Mindlessness: Social Responses to Computers*. *Journal of Social Issues*, 56(1), 81–103.
Nass, C., Moon, Y., & Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of Applied Social Psychology, 27*(10), 864–876.

Nissenbaum, H. (2010). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.
Noble, D. F. (1984). *Forces of Production: A Social History of Industrial Automation*. Knopf.
Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press.
Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology*, 20(1), 5–14.
Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., ... & Lazer, D. (2019). Machine behaviour. *Nature*, 568(7753), 477–486.
Reeves, B., & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press.
Rosenblat, A., & Stark, L. (2016). Algorithmic Labor and Information Asymmetries: A Case Study of Uber's Drivers. *International Journal of Communication*, 10, 3758–3784.
Rudman, L. A., & Glick, P. (2001). Prescriptive Gender Stereotypes and Backlash Toward Agentic Women. *Journal of Social Issues*, 57(4), 743–762.
Russell, S. J., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson. ISBN: 978-0-13-461099-3.
Selbst, A. D., & Barocas, S. (2018). The Intuitive Appeal of Explainable Machines. *Fordham Law Review*, 87(3), 1085–1139.
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). Proceedings of Machine Learning Research, 81. PMLR.
Shen, S., Zhang, C., & Wang, Y. (2021). Gender differences in trust and cooperation in human–AI interactions. *Journal of Artificial Intelligence Research, 70*, 1–20. https://doi.org/10.1613/jair.1.12230
Shin, D. (2021). *The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI*. *International Journal of Human–Computer Studies*, 146, 102551.
Sumi, R., Yasseri, T., Rung, A., Kornai, A., & Kertész, J. (2011). *Characterization and prediction of Wikipedia edit wars*. In *Proceedings of the 2011 ACM Web Science Conference (WebSci '11)* (pp. 1–3). ACM.
Tsvetkova, M., García-Gavilanes, R., Floridi, L., & Yasseri, T. (2017). *Even Good Bots Fight: The Case of Wikipedia*. *PLOS ONE*, 12(2), e0171774.
Tsvetkova, M., Yasseri, T., Meyer, E. T., Pickering, J. B., Engen, V., Walland, P., Lüders, M., Følstad, A., & Bravos, G. (2017). Understanding human–machine networks: A cross-disciplinary survey. *ACM Computing Surveys, 50*(1), 1–35.
Tsvetkova, M., Yasseri, T., Pescetelli, N., & Werner, T. (2024).A new sociology of humans and machines. *Nature Human Behaviour*, 8(10), 1864–1876.
Tufekci, Z. (2015). *Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency*. *Colorado Technology Law Journal*, 13(2), 203–218.
Vaccaro, M., Almaatouq, A., & Malone, T. W. (2024). *When combinations of humans and AI are useful: A systematic review and meta-analysis*. **Nature Human Behaviour**, *8*(12), 2293–2303.
Veale, M., & Borgesius, F. Z. (2021). Demystifying the draft EU Artificial Intelligence Act. *Computer Law Review International*, 22(4), 97–112.
West, M., Kraut, R., & Chew, H. E. (2019). *I'd Blush If I Could: Closing Gender Divides in Digital Skills Through Education.* UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000367416.
Whittaker, M., et al. (2018). *AI Now Report 2018*. AI Now Institute.
Wiener, N. (1950). *The Human Use of Human Beings: Cybernetics and Society*. Houghton Mifflin.
Wong, J., & Kim, J. (2023). *ChatGPT Is More Likely to Be Perceived as Male Than Female*. arXiv preprint arXiv:2305.12564.
Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). *Evidence for a Collective Intelligence Factor in the Performance of Human Groups*. *Science*, 330(6004), 686–688. https://doi.org/10.1126/science.1193147
Yasseri, T. (2024). Computational Sociology of Humans and Machines; Conflict and Collaboration. *arXiv preprint arXiv:2412.14606*
Yasseri, T. (2025, February 11). *The Memory Machine: How Large Language Models Shape Our Collective Past*. *Verfassungsblog*. https://verfassungsblog.de/the-memory-machine/
Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.