Understanding AI and Human Collaboration: What the Research Tells Us

Gabriela D. Dago

Intro: The AI Partnership - A Closer Look

Artificial intelligence (AI) is often compared to transformative General Purpose Technologies (GPTs) like electricity—innovations that did not merely enhance existing systems but redefined the underlying structure of society. Indeed, electricity's adoption required the reconfiguration of cities and the reinvention of industries.. This transition was neither immediate nor inevitable—it necessitated sustained investment in infrastructure, the establishment of regulatory standards, and even conceptions of energy, work, and time.

Today, AI presents a similar challenge. Its effective integration depends not only on advances in data, algorithms and computing power (Buchanan, 2020), but on the parallel development of frameworks that ensure ethical deployment, safety, cognitive resilience, and accountability. The gradual adoption of AI offers a critical window: to recognize and address the risk of AI systems subtly distorting human judgment, especially in tasks demanding expertise and critical thought.

Early experiences with AI highlight key risks. First, algorithmic limitations and contextual blind spots can degrade human performance, leading to inefficiencies, errors, or missed insights. Second, AI's speed and convenience can erode the user's inclination to validate sources, pursue deeper analysis, or build independent expertise. Left unchecked, habitual overreliance on AI could diminish human cognitive capabilities over time.

Acknowledging these risks—and deliberately designing for human-AI complementarity—is essential. This article examines recent research on the effects of suboptimal human-AI outcomes and explores strategies to cultivate stronger, more resilient collaboration models, aligning artificial computation with human judgment to maximize long-term potential.

I. AI's Impact on Task Performance: HR and Loan Application Case Studies

The increasing integration of AI across various sectors promises enhanced efficiency and reduced costs; however, it also introduces the pressing challenge of algorithmic bias, which can lead to worse outcomes, particularly in tasks requiring nuanced analysis. This section will explore this issue through the case studies of hiring practices and loan application processes, two domains where the impact of AI bias is particularly noticeable.

Recent evidence has highlighted the potential for underlying biases within AI tools, stemming from their training on flawed datasets. Algorithmic bias can manifest in several forms, each posing distinct risks to the fairness and accuracy of automated decision-making systems. Specifically, omitted variable bias occurs when relevant predictor variables are excluded from the model. This can lead to an incomplete model, or, more critically, to inaccurate or misleading results (Patel & Watts, 2025). Linking bias occurs when erroneous inferences are drawn about individuals based on the attributes of their network

connections, potentially misrepresenting individual characteristics (Schwartz et al., 2022).. Aggregation bias results from the inappropriate generalization of population-level trends to individual cases, thereby neglecting individual heterogeneity and undermining the suitability of models in diverse contexts (Mehrabi et al., 2022). These biases can materialize in various ways, often with significant real-world consequences.

In Human Resources, for example, the adoption of AI tools to streamline recruitment can inadvertently perpetuate existing biases. If an organization's demographics have been historically skewed, AI systems may learn to favor candidates with profiles similar to the existing employees, overlooking qualified individuals from underrepresented groups of women or minorities—AI models trained on such data may internalize and reproduce these inequities. This dynamic represents a form of measurement bias, wherein the evaluative metrics themselves are skewed toward the dominant group's characteristics (Albaroudi, Mansouri & Alameer, 2023). This phenomenon was evident in Amazon's hiring tool,which systematically favored male applicants because it was trained on a decade of predominantly male resumes (Dastin, 2018). This failure was not the result of overt programming choices but of passive data-driven learning from a biased historical corpus, demonstrating how seemingly neutral AI systems can entrench pre-existing disparities.

Similarly, in the financial sector, AI models used in credit evaluations risk perpetuating discriminatory practices under the guise of mathematical objectivity. Garcia, Garcia, and Rigobon (2024) observe that machine learning systems "consolidate any existing discrimination behavior under the veil of outcomes' precision and accuracy." In loan approval processes, models trained on historical lending data—already shaped by decades of biased practices—may inadvertently penalize applicants from marginalized groups. Again, this form of bias does not stem from malicious intent but from uncritical absorption of flawed historical patterns, allowing systemic inequities to persist under a new technological facade. Still, the absence of intent is no defense against the consequences. These biases undermine fairness, constrain organizational potential, and perpetuate structural inequalities across society.

II. The Cognitive Impact of AI Dependence

The previous section demonstrated how AI can entrench and amplify existing biases, producing unfair outcomes in domains such as hiring and lending. Yet the risks of AI extend beyond biased results. This section examines a more insidious threat: the erosion of essential human skills and critical thinking through over-reliance on automated systems. As AI becomes increasingly sophisticated and embedded in decision-making, users may develop an unwarranted trust in machine outputs — a phenomenon known as automation bias, extensively studied by Kahn, Probasco, and Kinoshita (2024) in AI Safety and Automation Bias. Automation bias impairs independent analysis and judgment, causing users to overlook or rationalize errors even when contradictory evidence is available. Over time, this dynamic fosters learned helplessness, as individuals begin to cede cognitive responsibility to AI and lose confidence in their own abilities. The result is a dangerous feedback loop: diminished human oversight combined with escalating system complexity.

Without deliberate intervention, this erosion of critical faculties will compromise the responsible integration of AI, increasing the likelihood of catastrophic errors and weakening human control precisely when it is most needed.

This erosion of human capabilities is not limited to the domain of analytical skills; it also touches upon broader cognitive functions. Gerlich (2025) and Valenzuela et al. (2024) raise concerns about the broader cognitive impact of AI dependence. Gerlich (2025) specifically investigates how the use of AI tools influences critical thinking skills, defining the related concept of cognitive offloading as the delegation of cognitive tasks to external aids like AI. His research reveals a significant negative correlation between frequent AI tool usage and critical thinking abilities, suggesting that as individuals increasingly rely on AI for tasks such as information retrieval and problem-solving, their own capacity for analysis, evaluation, and inference may weaken. This highlights a key concern: that AI's convenience and efficiency may discourage users from engaging in the active cognitive processes necessary to maintain and develop essential skills. Furthermore, the design of AI systems themselves can contribute to this over-reliance: user interfaces that are overly intuitive and seamless can create an illusion of infallibility, while 'black box' algorithms erode user understanding of the decision-making process, increasing dependence. Valenzuela et al. (2024) similarly caution against the potential for AI to 'de-skill' users, not only in task-oriented contexts but also by impacting emotional and social skills, further underscoring the multifaceted nature of this challenge.

III. The Root Causes: Digging Deeper

3.1 The Mystery of AI Decisions:

A major reason for over-reliance on AI is the lack of transparency in many systems. Users often struggle to understand the underlying logic of AI models, particularly those using deep learning and neural networks. The patterns within these networks are inherently complex and not fully explainable (Eschenbach 2021). This so-called 'black box' nature of AI, where the algorithmic processes are opaque, can erode user understanding and trust in their own judgment. Consequently, rational actors may default to the AI's output, irrespective of its potential fallibility.

3.2 Feeling Overwhelmed by Information:

AI systems, with their ability to process and present vast amounts of information, can also contribute to cognitive overload. In one study by Google Deep Mind, experts used AI assistants to rate AI models. During the rating process, they allowed the AI agents to aide in the deliberation phase by providing contrasting “evidence, reasoning and judgements”. In short, some AI agents operated or communicated in a debate-style. However, providing this information alone was not enough, and researchers point to the fact that it may have actually overwhelmed participants (Bridgers et al., 2024). The sheer volume of data and options provided by AI tools can strain human cognitive capacity, making it difficult for

users to focus, analyze information effectively, and make well-informed decisions. This information overload can paradoxically lead to a reliance on AI to simplify the decision-making process, even if that simplification comes at the cost of critical thinking.AI systems' capacity to process and present extensive information can induce cognitive overload. A Google DeepMind study involving experts rating AI models revealed that AI assistants providing contrasting evidence, reasoning, and judgments during deliberation, effectively operating in a debate-style, overwhelmed participants (Bridgers et al., 2024). This suggests that the sheer volume of data and options offered by AI tools can strain human cognitive resources, hindering focus, effective analysis, and informed decision-making. Paradoxically, this information overload may foster reliance on AI to simplify decision processes, potentially sacrificing critical thinking in the pursuit of simplification.

3.3 Trusting Too Much, Too Easily:

Another consequential risk is giving AI systems human-like authority and reliability. Molly Crockett, a cognitive psychologist and neuroscientist at Princeton University, cautions, "'These tools are being anthropomorphized and framed as humanlike and superhuman. We risk inappropriately extending trust to the information produced by AI'" (Scientific American, 2024). This 'automation bias' can lead users to overlook AI errors or limitations, even when there is evidence that the AI is incorrect. Factors such as the perceived authority of technology and the efficiency gains offered by AI can contribute to this over-trust, further exacerbating the risk of over-reliance and its associated negative consequences

IV. Conclusion: Navigating the Complexities of the AI Partnership

The preceding analysis reveals a complex interplay between the benefits and challenges of AI integration, particularly concerning its impact on human cognition and decision-making. AI offers undeniable potential for enhancing efficiency, productivity, and access to information. However, as this exploration has shown, the AI partnership is not without its inherent limitations and potential pitfalls.

As we have seen, several factors contribute to these challenges. The **lack of transparency** in many AI systems makes it difficult for users to understand the reasoning behind AI recommendations, fostering over-reliance and hindering effective collaboration. Moreover, the sheer **volume of readily available information** that AI can provide may overwhelm users, straining cognitive capacity and impeding any chance for AI-human collaboration. Moreover, **seamlessly designed AI interfaces** can quickly become entrenched in workflows, encouraging dependence. Finally, the human tendency to **over-trust automation** can lead to an uncritical acceptance of AI outputs.

To navigate these challenges and foster a more effective and responsible AI partnership, several strategies are essential.

Increased Transparency

Developing more transparent AI systems is crucial for building trust and accountability. Transparency in AI involves providing users with insight into how the system makes decisions, the data it uses, and the algorithms that underpin its processes. This can be achieved through clear documentation, open-source models, and user-friendly interfaces that allow non-experts to grasp the fundamental workings of AI. By understanding the "why" and "how" behind AI's conclusions and recommendations, users are more likely to trust the system, engage with it meaningfully, and hold it accountable when necessary. Transparency also ensures that biases in AI algorithms are more easily identified and corrected, which is essential for fostering fairness in automated decision-making.

Explainable AI (XAI)

Explainable AI (XAI) represents a critical evolution in AI technology, aiming to make machine learning models more interpretable to humans. With complex models, such as deep learning, it can be challenging for users to understand why a model arrived at a particular decision. XAI addresses this by providing tools that explain AI's reasoning in human-understandable terms. Techniques like feature importance analysis and decision visualization help users see which factors contributed most to a model's output and how those factors were weighed. For example, in a hiring algorithm, XAI can highlight the key traits (e.g., experience, education level) that influenced a candidate's recommendation. This level of explainability can reduce skepticism, increase trust, and enable users to better evaluate the decisions made by AI systems (Hofeditz, L., et. al, 2022).

Human-in-the-Loop (HITL)

The Human-in-the-Loop (HITL) approach emphasizes the importance of human oversight in AI decision-making. Rather than relying on fully autonomous AI systems, HITL ensures that human judgment remains integral to the process, particularly in complex or high-stakes situations. This approach combines the speed and scalability of AI with the nuanced understanding, ethical considerations, and emotional intelligence that humans bring. In practice, this could involve using AI to provide recommendations or data analysis, but requiring human decision-makers to review and approve key actions, such as loan approvals or medical diagnoses. By maintaining an active role for humans, HITL not only enhances the ethical alignment of AI systems but also mitigates the risks of over-reliance on automated decisions that might not fully account for context or human values.

Future Research Directions

As AI continues to evolve, future research should focus on several critical areas to ensure that AI systems continue to support human well-being and cognitive enhancement:

1. **Long-Term Cognitive Effects**: More research is needed to explore the long-term cognitive effects of interacting with AI tools. For example, how might reliance on AI for decision-making impact a person's long-term critical thinking skills, memory retention, or ability to solve problems independently? Understanding these effects will be vital for designing AI systems that preserve and enhance human cognitive capacities rather than diminish them over time.

2. **Critical Evaluation Skills**: In AI-driven environments, users must be equipped with the skills necessary to critically evaluate AI-generated outputs. Future research should investigate effective methods for teaching individuals how to assess AI recommendations, identify biases, and make informed decisions based on the outputs of AI systems. This could involve developing educational programs or tools that empower users to question, validate, and complement AI-generated insights with their own expertise.

3. **AI Systems for Cognitive Augmentation**: Finally, research should focus on developing AI systems that enhance human decision-making and cognitive abilities. These systems should be designed not as replacements for human intelligence but as tools that help users think more clearly, explore ideas more creatively, and arrive at better conclusions. AI should support rather than supplant human judgment, promoting collaboration rather than competition between humans and machines.

By focusing on these research areas, we can ensure that AI technologies evolve in a way that maximizes their potential for human enhancement, while minimizing the risks associated with over-dependence or bias. This approach will help to create a future where AI and humans work together in harmony, leveraging the strengths of both to solve complex problems and drive innovation.

References

Albaroudi, E., Mansouri, T., & Alameer, A. (2024). A comprehensive review of AI techniques for addressing algorithmic bias in job hiring. *AI, 5*(1), 383-404. https://doi.org/10.3390/ai5010019

Dastin, J. (2018, October 11). Insight—Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Available online: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-922showed-bias-against-women-idUSKCN1MK08G/

Crockett, M. (2023). The psychology of AI trust and anthropomorphism. *Journal of Artificial Intelligence Studies, 45*(2). Retrieved from https://www.journalofaistudies.org/articles/psychology-ai-trust

von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. *Philosophia Technologia, 34*, 1607-1622. https://doi.org/10.1007/s13347-021-00477-0

Bridgers, S., Jain, R., Greig, R., & Shah, R. (n.d.). *Human-AI complementarity: A goal for amplified oversight*. Retrieved from https://deepmindsafetyresearch.medium.com/human-ai-complementarity-a-goal-for-amplified-oversight-0ad8a44cae0a

Hofeditz, L., Clausen, S., Riess, A., Mirbabaie, M. I. (2022) *Applying XAI to an AI-based system for candidate management to mitigate bias and discrimination in hiring.* Retrieved from *https://www.researchgate.net/publication/366440619_Applying_XAI_to_an_AI-based_system_for_candidate_management_to_mitigate_bias_and_discrimination_in_hiring*

Garcia, A. C. B., Garcia, M. G. P., & Rigobon, R. (2024). Algorithmic discrimination in the credit domain: What do we know about it? *AI & Society, 39*, 2059-2098. https://doi.org/10.1007/s00146-023-01676-3

Buchanan, B. (2020, August). The AI triad and what it means for national security strategy. *Center for Security and Emerging Technology*. Retrieved from https://cset.georgetown.edu/publication/the-ai-triad-and-what-it-means-for-national-security-strategy/

McGovern Foundation. (2025, January 21). Bias in predictive machine learning: Part 2 - Data to algorithm bias. *The Patrick J. McGovern Foundation*. Retrieved from https://medium.com/patrick-j-mcgovern-foundation/bias-in-predictive-machine-learning-part-2-data-to-algorithm-bias-703aee5c9bc5

National Institute of Standards and Technology. (n.d.). *Special publication NIST.SP.1270*. Retrieved from https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022, January 25). A survey on bias and fairness in machine learning. *arXiv*. https://arxiv.org/pdf/1908.09635