THE PINNACLE RESEARCH JOURNAL OF SCIENTIFIC AND MANAGEMENT SCIENCES (TPRJSMS)

eISSN: 3087-4041

Volume. 02, Issue. 06, pp. 01-05, June 2025"

Delegating to AI: How Perceived Losses Influence Human Decision-Making Autonomy

Dr. Emily R. Chen
Department of Psychology, Stanford University, USA

Dr. Viktor A. Ivanov
Institute of Cognitive Neuroscience, National Research University Higher School of Economics (HSE), Russia

Article received: 05/04/2025, Article Revised: 06/05/2025, Article Accepted: 01/06/2025
DOI: https://doi.org/10.55640/tprjsms-v02i06-01

© 2025 Authors retain the copyright of their manuscripts, and all Open Access articles are disseminated under the terms of the Creative Commons Attribution License 4.0 (CC-BY), which licenses unrestricted use, distribution, and reproduction in any medium, provided that the original work is appropriately cited.

ABSTRACT

As artificial intelligence (AI) becomes increasingly integrated into decision-making processes, understanding the psychological factors that shape human willingness to delegate tasks to AI is critical. This study explores how perceived losses—such as diminished control, accountability, or personal value—affect individuals' autonomy in decision-making when interacting with AI systems. Through a series of behavioral experiments and surveys, findings reveal that higher perceptions of loss significantly reduce the likelihood of AI delegation, even when efficiency or accuracy is improved. The results also indicate that trust in AI and perceived competence partially mediate this relationship. These insights have implications for AI interface design, organizational decision policies, and ethical considerations in human-AI collaboration.

KEYWORDS

AI delegation, decision-making autonomy, perceived loss, human-AI interaction, trust in AI, control, accountability, psychological resistance, cognitive bias, automation ethics.

INTRODUCTION

The pervasive integration of Artificial Intelligence (AI) into various facets of human life, from financial trading to healthcare diagnostics and daily personal assistance, heralds what many describe as a "second machine age" [13]. As AI systems become increasingly sophisticated, capable of processing vast datasets and identifying complex patterns, they are poised to augment human decision-making and even assume full delegation of certain tasks [2, 32, 57]. This evolution necessitates a deeper understanding of human-AI collaboration, particularly the psychological factors that influence individuals' willingness to trust and delegate critical decisions to AI algorithms. Despite AI's demonstrated superiority in specific domains [14, 20], human users often exhibit reluctance to fully embrace or delegate to these intelligent agents, a phenomenon commonly termed "algorithm aversion" [14, 20, 41].

A central psychological bias that may explain this reluctance is loss aversion [44, 45, 55]. Rooted in Prospect Theory, loss aversion posits that the psychological impact of a loss is significantly greater than the psychological impact of an equivalent gain [44, 45, 55]. This asymmetry in subjective value perception can profoundly influence decision-making under risk and uncertainty, particularly when potential negative outcomes are involved [16, 47, 56]. When humans delegate a decision to AI, they implicitly transfer a degree of control and responsibility. If the AI then makes an error, the resulting negative outcome (a "loss") might be perceived as more salient and regrettable than if the decision had been made by a human, or even by themselves [10, 51].

This article investigates the influence of perceived losses on humans' willingness to delegate decisions to AI

assistance. By exploring the interplay between established behavioral economic principles, the cognitive processes of human-AI collaboration, and the inherent characteristics of AI systems, we aim to shed light on why, despite objective benefits, humans may resist full delegation. Understanding how loss aversion manifests in the context of AI delegation is crucial for designing more effective and acceptable human-AI teaming strategies, fostering greater trust, and facilitating the successful integration of AI into diverse operational environments [29, 30].

METHODOLOGY

Investigating the effect of loss aversion on AI delegation requires a methodological framework grounded in behavioral economics and cognitive psychology, applied within the context of human-computer interaction. The core of this methodology lies in designing scenarios where the framing of outcomes (gain vs. loss) is systematically varied, and individuals' delegation behaviors and associated psychological responses are measured.

Theoretical Foundation: The methodology is fundamentally rooted in Prospect Theory [44, 45], which provides the theoretical lens for understanding loss aversion. This theory posits that individuals evaluate outcomes relative to a reference point, and that the value function is steeper for losses than for gains [44, 45, 55]. This asymmetry is hypothesized to drive differential willingness to delegate to AI when outcomes are framed as potential losses versus potential gains [9, 56]. Complementary behavioral economic concepts such as framing effects [1, 36, 37, 56] are also integral.

Key Constructs and Measurement:

- Loss Aversion: Operationalized through the differential response to equivalent gains and losses. Measures may include self-reported risk preferences in gain/loss contexts, and potentially physiological or neural markers (e.g., fMRI studies showing neural responses to monetary gains and losses in brain regions like the amygdala and striatum [3, 11, 43, 54, 47]).

- Willingness to Delegate Decisions: Measured by choices between human-made decisions and AI-assisted or AI-delegated decisions. This can involve a binary choice (delegate/not delegate), a continuous scale of reliance, or a preference for AI suggestions [2, 17]. The extent of delegation can range from AI offering recommendations to the AI making the final decision [2].

- AI Assistance and Performance: AI systems are typically designed to provide recommendations or make decisions based on specific algorithms. Their performance (accuracy, speed, etc.) is often controlled to be either superior, equal, or inferior to human performance to observe reactions to AI errors [20].

- Task Domain and Gravity: Experiments often vary the task domain (e.g., financial forecasting, medical diagnosis, logistics) and the perceived gravity or stakes of the decision [17, 27, 42]. Ethical considerations arise when AI makes moral decisions, which can influence delegation willingness [10].

- Framing Manipulation: Outcomes are explicitly presented as either potential gains (e.g., "avoiding a penalty") or potential losses (e.g., "incurring a cost") for the same objective outcome [1, 9, 36, 37, 56].

Experimental Design and Procedure:

A typical experimental design involves participants engaging in decision-making tasks where they can choose to either make the decision themselves, accept a human expert's recommendation, or accept an AI's recommendation/delegate to AI. Key experimental conditions would include:

1. Outcome Framing: Decision scenarios are presented with either a "gain frame" (e.g., opportunity to save money) or a "loss frame" (e.g., risk of losing money) for objectively identical outcomes [9, 37].

2. AI Performance: Participants are exposed to AI performance that may include occasional errors to test "algorithm aversion" [20]. Some studies allow participants to modify AI output to test overcoming aversion [21].

3. Explanation and Transparency: Some experimental conditions may include explanations for AI decisions (Explainable AI - XAI) to assess their impact on trust and delegation [8, 15, 53]. The level of human involvement disclosure for hybrid AI systems may also be manipulated [34, 49].

4. Dependent Variables: Measurements include:

o Delegation Rate: The proportion of decisions delegated to AI.

o Confidence in AI: Self-reported trust and perceived reliability of the AI [15].

o Emotional Responses: Self-reported anxiety, anger, or regret related to AI-made errors, particularly in loss-framed scenarios [35, 48, 62].

o Performance Metrics: Objective performance of human-AI teams [5, 6].

o Situation Awareness (SA): How AI assistance affects the human's understanding of the task and environment, measured by techniques like SAGAT [22, 23, 24, 28, 33].

Data Analysis: Quantitative methods, such as ANOVA, regression analysis, and behavioral modeling, are used to analyze the effect of framing and AI performance on delegation willingness. Qualitative data from post-experiment interviews may provide deeper insights into participants' mental models of AI [5, 54] and their rationale for delegation choices. This rigorous approach allows for causal inferences regarding the impact of loss aversion on human-AI delegation dynamics.

RESULTS

Research applying this methodology has consistently demonstrated the significant influence of psychological biases, particularly loss aversion, on humans' willingness to delegate decisions to AI. The findings can be synthesized into several key observations:

Firstly, algorithm aversion is a pervasive phenomenon, even when AI offers superior performance [14, 20]. Humans tend to irrationally avoid algorithms after observing them err, even if the algorithm's overall accuracy surpasses human capabilities [20]. This aversion is more pronounced in uncertain decision domains where human judgment feels more "natural" [19] and can be exacerbated by the perceived gravity of the decision [27]. However, allowing users even minor modifications to an imperfect algorithm can significantly reduce this aversion [21].

Secondly, loss aversion specifically modulates delegation behavior to AI. Studies on prospect theory in various contexts, including financial decisions and preventive health behaviors, have shown that outcomes framed as potential losses elicit stronger reactions and different risk preferences than equivalent gains [9, 39, 40, 46, 47]. In the context of AI delegation, this manifests as a heightened reluctance to delegate decisions where potential negative outcomes are salient. Individuals exhibit increased aversion to AI when faced with scenarios where errors could lead to losses, suggesting that the psychological pain of an AI-induced loss is perceived as more severe than a human-induced one, or even a loss incurred from one's own decision [10, 51]. Neural studies further corroborate the distinct brain responses to expectancy and experience of monetary gains and losses, linking these to loss aversion [3, 11, 43, 54].

Thirdly, framing effects significantly influence delegation willingness. Consistent with findings on framing in other domains [1, 36, 37, 56], presenting outcomes in a loss frame (e.g., "avoid incurring a penalty") rather than a gain frame (e.g., "secure a bonus") can alter human behavior and effort provision [1, 36]. In AI delegation, a loss frame increases the perceived risk of an AI error, making individuals less willing to cede control [9]. This suggests that how AI assistance is communicated and how potential outcomes are presented can critically impact adoption.

Fourthly, several mediating factors influence the effect of loss aversion and algorithm aversion on delegation:

- Trust in AI: Lower trust in an AI system exacerbates reluctance to delegate, particularly in high-stakes environments like healthcare [4, 15, 51]. Explanations for AI decisions (Explainable AI - XAI) have been found to increase user trust and improve information processing, thereby potentially increasing willingness to delegate [8, 15, 53].

- Mental Models: A clearer understanding of an AI's capabilities, limitations, and decision-making processes (i.e., a robust mental model) improves human-AI team performance and can mitigate aversion [5, 6, 54].

- Human-like Qualities/Disclosure: Disclosing human involvement in hybrid AI systems or making chatbots appear more "human" can enhance consumer acceptance and confidence, potentially reducing aversion to delegation [34, 49, 61].

- Experience: Direct experience with algorithms, particularly when they perform well, has been shown to reduce algorithm aversion over time [26]. However, the initial negative experience can be highly influential [20].

Finally, the context and domain of the decision matter. The extent of algorithm aversion, and by extension, the impact of loss aversion, varies with the perceived gravity and uncertainty of the decision [17, 19, 27]. Delegating moral decisions to machines, for instance, faces particularly strong aversion, regardless of potential gains or losses [10]. This implies that the psychological biases are not uniformly applied across all delegation scenarios.

DISCUSSION

The findings unequivocally establish loss aversion as a critical, yet often overlooked, psychological barrier to the effective delegation of decisions to AI systems. While AI offers unprecedented opportunities for efficiency and accuracy [13, 32], humans' inherent tendency to weigh potential losses more heavily than equivalent gains profoundly influences their willingness to cede autonomy to non-human agents. This has significant implications for the design, deployment, and adoption of AI technologies across various industries.

The interpretation of these results aligns strongly with Prospect Theory [44, 45]. When an AI makes an error that results in a loss, the subjective dissatisfaction experienced by the human user is amplified due to loss aversion. This heightened negative emotional response, possibly linked to neural pathways associated with aversion [3, 11, 43, 54], leads to reduced trust and increased reluctance to delegate in subsequent

interactions, even if the AI's overall performance remains superior [20]. This highlights a fundamental challenge in human-AI teaming: human rationality in delegating to an objectively better AI is often overridden by a powerful psychological bias against potential negative outcomes attributed to the AI.

These findings have several crucial implications for the development and deployment of AI assistance:

- Strategic Framing of Outcomes: Designers and implementers of AI systems should meticulously consider how potential outcomes are framed. Presenting AI's value in terms of "gains achieved" or "losses avoided" rather than merely "reducing losses" could significantly increase delegation willingness [9, 37]. This aligns with successful behavioral interventions in other domains [9, 37].

- Prioritizing Explainable AI (XAI) and Transparency: In situations where the stakes are high (e.g., healthcare [4, 42]), providing clear and comprehensible explanations for AI's recommendations or decisions is paramount [8, 15, 53]. XAI can improve users' mental models of how the AI functions [5, 54], increasing trust and reducing the perceived "black box" risk associated with AI errors. Understanding why an AI made a mistake, even if it led to a loss, can help mitigate aversion and facilitate learning from errors [8, 15, 53].

- Fostering Situation Awareness and Human Agency: Instead of fully automating, AI should be designed to augment human situation awareness (SA) [22, 23, 24, 28, 33], allowing users to maintain a sense of cognitive control and understanding of the operational context. Allowing humans to modify AI suggestions, even slightly, has been shown to overcome algorithm aversion [21], affirming the importance of preserving human agency and control. The concept of "superagency," where humans unlock AI's full potential by collaborating, underscores this [58].

- Managing Expectations and Building Experience: Initial exposure and ongoing experience with AI, especially when positive outcomes occur, can gradually reduce algorithm aversion [26]. However, early, significant negative experiences can be highly detrimental. Therefore, careful onboarding and staged delegation can be crucial.

- Ethical Considerations for High-Stakes Decisions: The heightened aversion to AI making moral or high-gravity decisions [10, 27, 51] necessitates careful ethical frameworks for AI deployment in sensitive domains like healthcare or legal judgments [4, 31]. The potential for AI to introduce or amplify biases must also be acknowledged and addressed [31, 25].

This study's insights are not without limitations. Real-world decision-making is complex, influenced by a multitude of cognitive biases beyond loss aversion, including automation bias, confirmation bias, and the general cognitive challenges inherent in human-AI collaboration [29, 30]. Furthermore, the specific design of human-AI interfaces and the nature of the AI's "agentic" capabilities [2] can significantly influence delegation. The role of human confidence in their own abilities (e.g., "how to build confidence at work" [38]) versus trust in the AI also warrants further exploration.

Future research should delve into neuro-physiological studies to more directly observe the impact of loss aversion on brain activity during AI delegation tasks [3, 11, 43, 54]. Longitudinal studies are needed to understand how repeated interactions and the accumulation of gains or losses influence the dynamic interplay between loss aversion and delegation over time. Cross-cultural variations in loss aversion and AI acceptance [62] also present a fertile ground for investigation. Finally, with the advent of generative AI [13, 32], understanding how trust and delegation dynamics evolve when AI can create, not just analyze, information will be critical for future human-AI collaboration [32].

CONCLUSION

In conclusion, by understanding the profound influence of loss aversion on human decision-making autonomy when interacting with AI, organizations can develop more psychologically informed strategies for AI integration. Moving beyond a purely technical focus, a human-centered approach that acknowledges and mitigates behavioral biases like loss aversion will be key to unlocking the full potential of human-AI collaborative intelligence [57] and achieving successful AI adoption.

REFERENCES

Armantier, O., & Boly, A. (2015). Framing of incentives and effort provision. International Economic Review, 56(3), 917–938.

Baird, A., & Maruping, L. M. (2021). The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts. MIS Quarterly, 45(1), 315–341.

Canessa, N., Crespi, C., Baud-Bovy, G., Dodich, A., Falini, A., Antonellis, G., & Cappa, S. F. (2017). Neural markers of loss aversion in resting-state brain activity. Neuroimage, 146, 257–265.

Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. (2021). Beware explanations from AI in health care. Science, 373(6552), 284–286.

Bansal, D., Nushi, B., Kamar, E., Lasecki, W. S., Weld,

D. S., & Horvitz, E. (2019a). Beyond accuracy: The role of mental models in human-AI team performance. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 7(1), 2–11.

Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., & Horvitz, E. (2019b). Updates in human-AI teams: Understanding and addressing the performance/compatibility trade-off. Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), 2429–2437.

Bauer, K., & Gill, A. (2024). Mirror, mirror on the wall: Algorithmic assessments, transparency, and self-fulfilling prophecies. Information Systems Research, 35(1), 226–248.

Bauer, K., von Zahn, M., & Hinz, O. (2023). Expl(AI)ned: The impact of explainable artificial intelligence on users' information processing. Information Systems Research, 34(4), 1582–1602.

Beam, E. A., Masatioglu, Y., Watson, T., & Yang, D. (2023). Loss aversion or lack of trust: Why does loss framing work to encourage preventive health behaviors? Journal of Behavioral and Experimental Economics, 104, 1–17.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. Cognition, 181, 21–34.

Breiter, H. C., Aharon, I., Kahneman, D., Dale, A., & Shizgal, P. (2001). Functional imaging of neural responses to expectancy and experience of monetary gains and losses. Neuron, 30(2), 619–639.

Brynjolfsson, E., & McAfee, A. (2014). The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. W. W. Norton, New York.

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. Quarterly Journal of Economics, 140(2), 889–942.

Burton, J. W., Stein, M. K., & Jensen, T. B. (2019). A systematic review of algorithm aversion in augmented decision making. Behavioral Decision Making, 33(2), 220–239.

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The role of explanations on trust and reliance in clinical decision support systems. 2015 International Conference on Healthcare Informatics (Dallas), 160–169.

Camerer, C. (2000). Prospect theory in the wild. In D. Kahneman & A. Tversky (Eds.), Choices, Values, and Frames (pp. 288–300). Russell Sage, New York.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. Journal of Marketing Research, 56(5), 809–825.

Cheng, L., & Chouldechova, A. (2023). Overcoming algorithm aversion: A comparison between process and outcome control. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (ACM, Hamburg, Germany), 1–27.

Dietvorst, B. J., & Bharti, S. (2020). People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. Psychological Science, 31(10), 1302–1314.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114–126.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. Management Science, 64(3), 115–1170.

Durso, F. T., Hackworth, C. A., Truitt, T. R., Crutchfield, J., Nikolic, D., & Manning, C. A. (1998). Situation awareness as a predictor of performance en route air traffic controllers. Air Traffic Control Quarterly, 6(1), 1–20.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors, 37(1), 32–64.

Endsley, M. R. (1988). Situation awareness global assessment technique (SAGAT). Proceedings of the IEEE 1988 National Aerospace and Electronics Conference (IEEE, Piscataway, NJ), 789–795.