UNIVERSITÄT PADERBORN

HEINZ NIXDORF INSTITUT
UNIVERSITÄT PADERBORN

Kumulative Dissertation zum Themengebiet

Behavioral Effects in Human-Machine and Human-Human Interactions:

Economic Experiments on Dishonesty, Advice-Taking, and Overconfidence

Fakultät für Wirtschaftswissenschaften der

Universität Paderborn

zur Erlangung des akademischen Grades

Doktor der Wirtschaftswissenschaften

- Doctor rerum politicarum -

vorgelegte Dissertation von

Marius Protte

geboren am 16.04.1995

in Salzkotten

2025

Danksagung

Alle paar Jahre veröffentlich das Deutsche Zentrum für Hochschul- und Wissenschaftsforschung (DZHW) den sogenannten "Bildungstrichter" – eine Grafik, die den Bildungsverlauf von Kindern aus Akademiker- und Nicht-Akademiker-Haushalten im Hinblick auf die Anteile der jeweiligen Gruppe, die einen bestimmten Bildungsabschluss erlangen, nebeneinanderstellt. Diese Grafik liefert seit Jahrzehnten die gleiche zentrale Aussage: Ab Beginn eines Studiums unterscheiden sich die bedingten Häufigkeiten für den Erwerb des jeweils nächsthöheren akademischen Abschlusses zwischen den Kohorten nur unwesentlich. Dagegen verengt sich der Trichter auf Seiten der Kinder aus Nicht-Akademiker-Haushalten in der ersten Stufe – dem Anteil der Grundschüler, die irgendwann mal ein Studium beginnen – drastisch (in der letzten Erhebung 25 aus 100), und damit auf weniger als ein Drittel des Anteils von Akademiker-Kindern (78 aus 100). Im Schnitt schließen 1-2 von 100 Nicht-Akademiker-Kindern eine Promotion ab. Die Frage, ob Kinder es überhaupt einmal selbst in der Hand haben werden, diesen Trichter eines Tages durchzuspielen, wird also maßgeblich von anderen beantwortet...

Für Muddi.

Contents

Paper of the Dissertation

- **Protte, M., and Djawadi, B. D. (2025)**: Human vs. Algorithmic Auditors: The Impact of Entity Type and Ambiguity on Human Dishonesty. *Frontiers in Behavioral Economics*, 4, 1645749.

- **Lebedeva, A., Protte, M., van Straaten, D., and Fahr, R. (2024)**: Does the involvement of domain experts in the AI training affect their AI perception and AI adherence? An experiment in the context of industrial AutoML applications. *Advances in Information and Communication: Proceedings of the 2024 Future of Information and Communication Conference (FICC)*, 178-204. Springer Nature.

- **Protte, M., Fahr, R., and Quevedo, D. E. (2020)**: Behavioral Economics for Human-in-the-loop Control Systems design: Overconfidence and the hot hand fallacy. *IEEE Control Systems Magazine*, 40(6), 57-76.

- **Protte, M. (2025)**: Explaining apparently inaccurate self-assessments of relative performance: A replication and adaptation of Hoelzl & Rustichini (2005). Working Paper. https://doi.org/10.48550/arXiv.2507.15568.

Introduction and Synopsis

As the title suggests, this dissertation synthesizes insights from behavioral economics with contemporary challenges in human–machine interaction. Adopting an interdisciplinary perspective, it integrates both established and newly generated findings on human decision-making (see Section 1.1) – such as dishonesty, perceptual biases, and overconfidence – with emerging research questions related to real-world technological developments, particularly in the context of algorithmic and AI systems (see Section 1.2).

Employing controlled economic experiments (see Section 1.3), the dissertation systematically observes and evaluates human behavior toward both algorithmic and human counterparts, identifying behavioral patterns and potential inefficiencies. In doing so, it contributes to the expanding field of human–machine interaction – central to the megatrend of Industry 4.0 – by focusing on user perception, trust, and adherence to algorithmic information. Complementing this, it revisits foundational paradigms of behavioral economics by replicating and modifying a seminal study on overconfidence by [Hoelzl and Rustichini, 2005] nearly two decades after its original publication. These insights offer value not only for advancing the academic understanding of human behavior through experimental research, but also for informing practical applications in increasingly automated environments.

Within the broad domains of behavioral economics and human–machine interaction, the four distinct experimental studies presented in the following chapters shed light on how individuals engage in dishonest behavior in interactions with machines and humans, respond to algorithmic advice and feedback in cyber-physical systems, and assess their own abilities relative to others. Before introducing these studies, the following sections outline the contextual framework within which the experimental studies were conducted.

1.1 "I Tried So Hard and Got So Far": Bounded Rationality and Non-Monetary Preferences in Human Decision-Making

While classical economic models typically depict human decision-makers as perfectly rational agents who make optimal choices by processing all available information in accordance with formal logic and statistical rules, many real-world decision-making phenomena cannot be adequately explained by the principles and assumptions of neoclassical economic models [Camerer, 1999; Todd and Gigerenzer, 2003]. Instead, empirical findings repeatedly demonstrate that individuals operate within the constraints of bounded rationality – a concept originally introduced by Herbert A. Simon who posed the question: *"How do human beings reason when the conditions for rationality postulated by the model of neoclassical economics are not met?"* [Simon, 1989, p. 377].

Bounded rationality refers to the notion that human decision-making is constrained by factors such as restricted access to information, limited cognitive processing capacity, and time pressure [Simon, 1955, 1972]. As a consequence, individuals rely on mental shortcuts – known as *heuristics* – to navigate complexity in decision-making processes under risk uncertainty [Camerer and Loewenstein, 2004; Kahneman, 2003; Gigerenzer and Todd, 1999]. These heuristics enable individuals to find "satisficing" [Simon, 1972] rather than optimal solutions, which reflects adaptive behavior within environmental and cognitive constraints [Simon, 1955; Goldstein and Gigerenzer, 2002].

In the field of behavioral economics, two major perspectives on bounded rationality have evolved, each offering distinct interpretations of the role and impact of heuristics. On the one side, Daniel Kahneman and Amos Tversky emphasize the systematic errors – *"behavioral biases"* – that arise from heuristic reasoning. They show that such biases produce predictable patterns of judgment and decision errors, which can result in economic inefficiencies, while pointing out that even experts violate the normative standards of expected utility theory (EUT), and that financial do not necessarily mitigate these cognitive distortions [Tversky and Kahneman, 1974, 1981; Kahneman and Tversky, 1972, 1973; Tversky and Kahneman, 1973].

On the other side, Gerd Gigerenzer and colleagues argue that heuristics should not be regarded merely as cognitive flaws but as adaptive tools to function efficiently in real-world environments. According to this view, humans possess an *"adaptive toolbox"* of *"fast and frugal"* heuristics suited to specific decision types and environmental contexts [Gigerenzer and Todd, 1999]. Though not universally optimal, these heuristics can lead to outcomes that are efficient in terms of time, effort, and cognitive resources and, under certain conditions, may even outperform more complex decision strategies [Gigerenzer, 2001; Gigerenzer and Todd, 1999; Goldstein and Gigerenzer, 2002; Borges et al., 1999][1].

Crucially, bounded rationality should not be equated with irrationality [Selten, 2001]. While irrational behavior implies a complete lack of structure or purpose, bounded rationality refers to purposeful decision-making within the limitations of human cognition, i.e., behavior that may deviate from idealized models but not from subjective logic given cognitive constraints[Simon, 1978; March, 1978; Simon, 1986; Selten, 1998].

In summary, bounded rationality offers a more realistic foundation for understanding and modeling

[1]The debate between these two perspectives has generated a vibrant discourse. While Kahneman and Tversky emphasize the systematic errors associated with heuristic thinking [Tversky and Kahneman, 1974, 1983; Kahneman and Tversky, 1996], Gigerenzer contends that such conclusions often stem from artificially narrow experimental setups and misunderstandings of how people interpret probabilistic tasks [Gigerenzer, 1996; Hertwig and Gigerenzer, 1999; Goldstein and Gigerenzer, 2002]. He argues that the cognitive errors observed by Kahneman and Tversky often diminish or disappear when problems are framed using natural frequencies rather than abstract probabilities [Gigerenzer, 1991; Gigerenzer and Hoffrage, 1995; Hoffrage et al., 2000]. Nevertheless, both camps acknowledge that heuristics have their value – Kahneman and Tversky concede their general usefulness [Kahneman and Tversky, 1996], and Gigerenzer admits that heuristics can lead to biases in inappropriate contexts [Gigerenzer, 1996].

human decision-making. Although the term encompasses diverse theoretical interpretations, bounded rationality can be considered a cornerstone concept in behavioral research. While a universally integrated theory is still lacking, ongoing interdisciplinary efforts to study how people actually behave, in contrast to how they are prescribed to behave, continue to advance our theoretical and practical understanding. These insights not only improve explanation and prediction of human behavior but also form the basis for designing policies that accommodate rather than ignore human limitations [Gigerenzer, 2016].

Complementarily, human decision-making is also heavily shaped by non-monetary preferences, which often leads their behavior to diverge from rational optimization models such as expected utility theory [von Neumann and Morgenstern, 1944] in their own right. While bounded rationality refers to cognitive limitations that prevent individuals from acting in economically optimal ways, non-monetary preferences reflect additional motives for decision-making that extend beyond financial considerations. Behavioral economics research has documented a wide array of such preferences – including fairness, reciprocity, identity, social norms, moral values, and self-image – which can cause substantial deviations from traditional payoff- and utility-maximization assumptions of neoclassical economics.

Central to this field is the literature on *social preferences*, which highlights that individuals care not only about their own outcomes but also about the outcomes of others. This is exemplified in models of *inequity aversion*, which assume that people derive disutility from unequal distributions of income or resources [Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000]. For instance, in ultimatum games, individuals frequently reject offers perceived as unfair, even at a cost to themselves, contradicting the payoff-maximizing prediction that any non-zero offer should be accepted [Güth et al., 1982]. Similarly, humans often exhibit *altruism*, the willingness to help others even at personal cost. Altruistic behavior contradicts the standard economic assumption of self-interest but has been repeatedly observed in experimental and real-world contexts, such as charitable giving, blood donation, and volunteerism [Eckel and Grossman, 1996; Andreoni, 1990; Batson and Shaw, 1991]. *Fairness* considerations, moreover, extend beyond the principle of equity in material distribution and often encompass procedural justice, intentions, and perceived legitimacy of actions [Rabin, 1993; Konow, 2003]. For example, individuals may reject outcomes they view as unfair not only because of unequal payoffs but also due to perceived violations of moral or normative expectations [Cappelen et al., 2007]. All these behavioral patterns are closely linked to *reciprocity* – the tendency to respond kindly to kind actions and punishing unkind ones – which has been shown to underlie behaviors in various strategic settings, from labor markets to public goods provision [Rabin, 1993; Fehr and Gächter, 2000; Falk and Fischbacher, 2006].

Furthermore, individuals often act in ways that affirm their social *identity* or conform to group *norms*, even when doing so contradicts (material) self-interest [Akerlof and Kranton, 2000; Fehr and

Fischbacher, 2004; Duffy and Ochs, 2009; Bicchieri and Xiao, 2009]. This explains why people may choose jobs, consumption patterns, or political positions that align with their perceived identity or comply with group or societal norms rather than purely economic gain. *Moral* preferences and *ethical* considerations play a pivotal role. Individuals often make decisions based on internalized moral norms or ethical standards, even in anonymous and incentivized contexts. For example, many people refrain from lying or cheating in experiments despite having clear financial incentives to do so and not facing any risk of punishment [Abeler et al., 2019; Fischbacher and Föllmi-Heusi, 2013; Gneezy et al., 2018]. These behaviors suggest the presence of internal preferences for honesty, self-concept maintenance, or moral integrity, which lead to systematic deviations from payoff-maximization predictions. The same applies for preferences regarding risk and uncertainty. Contrary to expected utility theory – which assumes stable, context-independent risk neutrality – empirical research reveals that individual risk-taking behavior and ambiguity attitudes are often inconsistent and context-sensitive. They are influenced by factors such as emotions, perceived control, framing, and moral context, with most individuals tending toward risk and ambiguity aversion [Kahneman and Tversky, 1979; Holt and Laury, 2002; Loewenstein et al., 2001; Ellsberg, 1961; Gneezy et al., 2015].

These diverse decision rationales challenge the notion that behavior can be adequately explained by a utility functions based solely on monetary outcomes. Instead, non-monetary preferences introduce multiple, sometimes competing, goals and motivations that can lead to behavior that appears inconsistent or suboptimal from a traditional economic perspective, especially as individual preferences vary across demographic, cultural, and cognitive dimensions [Falk et al., 2018]. Although such behavior is not *payoff*-maximizing, it may still be *utility*-maximizing and thus subjectively rational from an individual's perspective [Andreoni and Miller, 2002; Bénabou and Tirole, 2011], circling back to Simons's thoughts on bounded rationality. Importantly, as with bounded rationality, the existence of non-monetary preferences does not imply irrationality. Preferences are not irrational [Benoît et al., 2022], they rather reflect an expanded conception of rational agency – one that recognizes humans as socially embedded actors with plural and context-dependent motivations.

In any case, it has become increasingly clear that human decision-making cannot be reduced purely to optimization or payoff-maximizing processes. Behavioral economics has identified a range of phenomena, artifacts, and stylized facts [Hirschman, 2016] and resulting economic outcomes based on, or related to, the interplay of bounded rationality and diverse preferences. These findings demonstrate that preferences and perceptions frequently deviate from traditional rational choice assumptions by violating assumptions of consistency, transitivity, or payoff dominance, reflecting the complex, value-laden nature of human decision-making.

This underscores the importance of studying human decision-making for both academic research

and practical application. In academic contexts, understanding the cognitive constraints and non-monetary preferences that shape individual choices helps refine economic theory to better reflect actual behavior rather than idealized prescriptions [Simon, 1955, 1972]. In practice, such insights are crucial for designing more effective strategies in public policy, healthcare, education, finance, law, and emerging technologies (see Section 1.2), as well as for improving predictions of market outcomes, voting behavior, consumer choices, and risk assessments [Jones, 1999; Hoffrage et al., 2000; Falk et al., 2018; Shiller, 2003]. Misjudging cognitive boundaries or neglecting non-monetary preferences can result in poor regulatory design or the failure of interventions that assume levels of rationality that humans do not possess or behavioral patterns to which they do not adhere. This is one why economists and policymakers increasingly design behaviorally informed interventions – such as "nudges" – that leverage social norms, moral cues, or identity salience to steer behavior in a certain direction in domains ranging from health and financial planning, to climate change [Thaler and Sunstein, 2008; Bicchieri, 2006; Bénabou and Tirole, 2011; Kamenica, 2012]. The practical relevance of behavioral insights for public policymaking has been further underscored by political leaders' growing engagement with behavioral science and the implementation of dedicated expert bodies. Notable examples include the establishment of the "Social and Behavioral Sciences Team" during the Obama administration [Congdon and Shankar, 2015], the launch of the European Commission's "Competence Centre on Behavioural Insights" [European Commission, 2019], and the World Health Organization's recent formation of the "Technical Advisory Group on Behavioural Sciences for Better Health" [World Health Organization, 2025].

Given this context, this dissertation aims to contribute to the understanding of human behavior in contexts of dishonesty (Chapter 2), reactions to advice and feedback (Chapters 3 and 4), and self-assessment (Chapter 5). It accounts for behavioral phenomena attributed to both cognitive limitations (Chapters 4 and 5, partly 3) and non-monetary preferences (Chapters 2 and 3, partly 5).

1.2 "O Brave New World": Human Behavior Toward Algorithmic Technologies

As artificial intelligence (AI), machine learning (ML), and algorithmic systems have become increasingly embedded in everyday life in recent years, human interactions with these technologies have garnered growing academic interest[2]. Once confined to the realm of science fiction, intelligent systems now influence decision-making in healthcare, finance, education, justice, public administration and a

[2]For simplicity, the terms artificial intelligence, machine learning, and algorithmic systems will largely be used interchangeably throughout this dissertation, as the focus lies not on the technical distinctions between these technologies, but rather on the human behavioral responses they elicit in decision-making contexts. Additionally, "machine" will be used as an umbrella term to refer to these systems collectively.

wide range of other domains [see e.g. Cheng et al., 2016; Gruber, 2019; Tao et al., 2021; Black et al., 2022; Zhai et al., 2021; Kleinberg et al., 2018; Kouziokasa, 2017; Bignami, 2022]. Understanding human behavior toward these intelligent machine systems – ranging from trust, cooperation, and reliance to fear, resistance, and overreliance [Parasuraman and Riley, 1997; Jussupow et al., 2024; Mahmud et al., 2022] – is therefore critical to their design, implementation, and regulation[3]. This interplay has profound implications for both researchers studying human-machine interaction and practitioners implementing and using these tools in real-world contexts.

As of June 2025, ChatGPT alone boasts approximately 800 million weekly active users (120–140 million daily) and handles over 1 billion interactions per day, ranking it among the world's five most visited websites with an estimated 5.5 billion unique monthly visits in May 2025 [Duarte, 2025; Semrush.com, 2025; Kaiser, 2025]. Meanwhile, from a business perspective, estimates by McKinsey & Company [2023] suggest that generative AI alone could create USD 2.6 to 4.4 trillion in economic value across industries, with investments in AI being found to be associated with both higher growth in sales and higher market valuations through product innovation [Babina et al., 2024][4]. In the first quarter of 2025, 210 S&P 500 companies (42%) mentioned "AI" in their earnings calls – well above the 5-year average of 114 – marking five consecutive quarters with 200+ mentions [Butters, 2025]. This widespread adoption – whether driven by genuine economic efficiency or merely by its appeal as a fashionable selling point – puts AI on the verge of shifting from a differentiator and marker of competitive success to a "hygiene factor", i.e., a basic expectation, in both product design and institutional strategy.

Similarly, a clear upward trend in the integration of algorithmic systems into economics research is evident – encompassing both their use as research tools [see e.g., Korinek, 2023; Christou, 2023; Khalifa and Albadawy, 2024] and their emergence as a research topic in their own right. Between 2013 and 2021, 6,949 AI-related publications across various economics sub-fields were recorded in the *EconPapers* database[5] – constituting 0.95% of all economics outputs of this time span – with the highest concentrations in econometrics and financial economics [Bickley et al., 2022]. A bibliometric analysis of the broader social sciences by Prieto-Gutierrez et al. [2023] shows an even stronger surge: by 2022, nearly 20,000 AI-related articles had been published[6], predominately in the fields of law, education, economics or ethics, with 85% of them appearing since 2008. Similar results are obtained by Hajkowicz

[3]The effects of the machine's physical embodiment – such as in the case of (anthropomorphic) robots [see e.g. Eyssel and Kuchenbrandt, 2012; Canning et al., 2014; Ullman et al., 2014; Sandoval et al., 2020; Petisca et al., 2022] – are beyond the scope of this dissertation.

[4]Meanwhile, the effect of AI on labor market outcomes, such as employment levels, productivity, skill demand, wages, income, and job vacancies, remains less clear [Carbonara et al., 2025].

[5]Publications – including journal articles, working paper, books and book chapters – are classified as such if they featured a JEL code [Bickley et al., 2022].

[6]The figure refers to articles classified under "Social Sciences" in the *Scopus* database that include the term "artificial intelligence" in the title, abstract, or keywords [Prieto-Gutierrez et al., 2023].

et al. [2023], using an even larger scope: 3.1 million of the 137 million (2.3%) peer-reviewed research publications captured in *The Lens* database during the period from 1960 to 2021 are AI-related[7], with the recent increase across practically all research fields of research (physical science, natural science, life science, social science, arts and humanities) being considered historically unprecedented in speed and magnitude and its implications expected to be impactful and long-lasting.

Meanwhile, despite their increasing ubiquity, substantial heterogeneity can be observed in the efficacy of human–machine interactions, with a collaboration between entities not necessarily outperforming either acting alone [Vaccaro et al., 2024]. While algorithmic reasoning – as deterministic, stepwise decision-making processes [Dietvorst and Bharti, 2020] – is characterized by reliability, consistency, and objectivity, these qualities cannot be equally expected from human agents, as highlighted in Section 1.1. Camerer [2019, p. 587] aptly illustrates this point by stating that "some common limits on human prediction might be understood as [...] poor implementations of machine learning".

Correspondingly, empirical and experimental research on human-machine interaction documents behavioral phenomena such as algorithm aversion and algorithm appreciation [Jussupow et al., 2024; Mahmud et al., 2022]. However, human perception of machine systems typically represents more of a fluid spectrum than a strict binary classification, as it typically varies depending on contextual factors such as the anticipated efficacy of, and the degree of trust placed in, a given algorithmic system [Fenneman et al., 2021]. This underscores the complexity and heterogeneity of human attitudes toward machine(-supported) decision-making. Even prior to the rise of digital systems, it has been pointed out that human decision-makers perceive and evaluate statistical algorithms differently from human judgment [Dawes et al., 1989] – despite the former demonstrating equal or superior performance – a pattern now also found in responses to AI and other digital systems [Dietvorst et al., 2015; Logg et al., 2019]. This discrepancy may be rooted in psychological, technological, or ethical origins [Dawes, 1979].

Beyond the biological baseline of being "flesh-and-blood", human interaction partners are characterized by free will and the associated discretionary power to apply or interpret given rules. This allows their decisions to be affected by psychological processes, such as empathy, moral sensibility, or strategic interest. In contrast, machines interaction partners rigorously apply rules with near perfect precision and consistency, which may trigger inefficient skepticism and disuse in case a machine error is witnessed [Jauernig et al., 2022; Dzindolet et al., 2002; Haslam, 2006; Dietvorst et al., 2015].

Furthermore, transparency plays a central role in this dynamic. Many AI systems – especially deep learning models – are perceived as "black boxes", producing outputs based on partially known or opaque inputs [Burrell, 2016; Tschider, 2020]. This opacity hinders user evaluation of reliability

[7] A publication is considered AI-related if it contains at least one of 214 phrases identified by expert working groups at the Organisation for Economic Cooperation and Development (OECD) in the title, abstract, or keywords [Hajkowicz et al., 2023].

and unbiasedness, fostering either undue skepticism [Dietvorst et al., 2015] or blind trust [Klingbeil et al., 2024]. This issue sparks discussion about explainability and interpretability of AI and ML [see e.g. Doshi-Velez and Kim, 2017; Burkart and Huber, 2021], as these processes are likely affected by humans' prior beliefs, cognitive biases, and social framing in their own right [Miller, 2019].

Moreover, besides philosophical and legal questions regarding the responsibility for potentially harmful decisions made by autonomous algorithmic systems in cases like self-driving cars [see e.g., Coeckelbergh, 2020; Mittelstadt et al., 2016], ethical considerations strongly shape human perceptions of algorithmic systems [Irlenbusch and Köbis, 2025] – especially toward aversion. In morally salient domains such as healthcare, criminal justice, and military – where decisions can carry serious consequences and impose externalities on third parties – algorithmic input is frequently rejected, even when it aligns with human judgment and yield efficient outcomes [Bigman and Gray, 2018; Gogoll and Uhl, 2018; Longoni et al., 2019]. This may reflect a sense of dehumanization through neglect of human uniqueness [Haslam, 2006], which may be mitigated by reintroducing a human agent who receives algorithmic advice rather than being replaced by an autonomous machine. In this setup, the human remains the final decision-maker who is supported – but not substituted – by the system [Bigman and Gray, 2018; Longoni et al., 2019]. However, advisory systems are not without risks, as algorithmic advisors can facilitate unethical behavior while not necessarily promoting ethical behavior [Leib et al., 2024; Köbis et al., 2021].

Together, these trends signal a profound shift not just technologically, but in how humans engage with intelligent systems and how economics as a discipline must adapt to study this new landscape. In sum, the growing ubiquity of human–machine interaction has transformed decision-making across domains, introducing new psychological, ethical, and epistemological challenges. As AI becomes an integral component of modern life, understanding how humans perceive, interact with, and are influenced by these systems is of increasing significance. Human behavior shapes not only how machines are used, but also how they are developed, governed, and evaluated. Misalignment between technological capabilities and human expectations may lead to dysfunction, harm, or ethical failure.

Research and practice must, therefore, explicitly account for the human component, its characteristics, behaviors, and limitations, when modeling human-machine interactions and designing interfaces, policies, and automation strategies. Such considerations are essential to avoid human users becoming the "weak link" in human-machine interactions, potentially compromising technological progress.

This dissertation addresses this challenge by examining human behavior toward machines and automation across three domains: dishonesty (Chapter 2), involvement in system training (Chapter 3), and in-the-loop decision-making (Chapter 4), before concluding with an investigation on self-assessment in human–human comparison (Chapter 5).

1.3 Economic Experiments for Studying Human Behavior

Methodologically, behavioral economics research has been inextricably intertwined with experimental approaches, as "experimental control is exceptionally helpful for distinguishing behavioral explanations from standard ones" [Camerer and Loewenstein, 2003, p. 6]. Such discrepancies between theoretically prescribed behavior and actual human decision-making ("behavioral messiness") pose a profound challenge for organizational leaders and policymakers alike. Hence, experiments serve as an important tool for bridging this gap between economic theory and real-world institutional design, thus generating practical value from economic science [Bolton and Ockenfels, 2012].

"If you want to understand a person, don't listen to his words. Observe his behavior." is a quote commonly attributed to Albert Einstein, although no record of it is found in his papers, letters, speeches, or interviews.

In this spirit, all four studies of this dissertation employ economic experiments – conducted either in the laboratory (Chapters 2 and 3) or online (Chapters 4 and 5) – as controlled and replicable settings for studying human decision-making. The observational data generated from these experiments provides the basis for quantitative statistical analysis (non-parametric and parametric) to identify systematic behavioral patterns [Croson, 2005].

Economic experiments aim to parallel real-world decision-making scenarios, while isolating specific variables of interest from complex environmental contexts in order to enable causal inference [Falk and Heckman, 2022; Pearl, 2003; Friedman and Sunder, 1994]. A defining feature of these experiments is the use of salient, decision-dependent monetary incentives, which are implemented to tie participants' choices to real personal – or, in some cases, collective – consequences. This approach addresses issues such as intention–behavior gaps [Carrington et al., 2010] and social desirability bias in self-reports [Nederhof, 1985], which may limit the validity of unincentivized studies or those that merely compensate participants for their time through fixed payments. Without meaningful salient incentives, results of surveys, interviews, or hypothetical scenario-based designs are unlikely to yield informative insights for research questions in multiple domains such as moral dilemmas, risk-taking, or prosocial behavior [Croson, 2005; Camerer and Hogarth, 1999; Smith, 1976].

Economic experiments further function as a "wind-tunnel" for practice, allowing researchers to test behavioral responses to institutional interventions in a controlled environment before implementing them in the real world [Bolton and Ockenfels, 2012] ("behavioral economic engineering"[8]), as real-world testing is often costly (as in the context of Chapter 2) or entirely unfeasible due to operational disruptions, risk of economic harm, or even the potential for physical danger to humans involved (as

[8]Prominent examples include the implementation of optimized retirement savings plans [Thaler and Benartzi, 2004] and the redesign of matching algorithms for medical residency placements in the United States [Roth and Peranson, 1999].

could be expected in Chapters 3 and 4). Therefore, despite typically varying in their external validity, depending on their alignment with real-world decision contexts, economic experiments often represent the closest approximation available, yielding valuable and cost-efficient insights. When properly designed, they usually provide strong internal validity, allowing for reliable inference about the effects of isolated variables and serving as a useful complement to other empirical research methods [Friedman and Sunder, 1994; Guala, 2005; Croson, 2005; Falk and Heckman, 2022].

1.4 Synopsis of the Individual Research Papers

While all chapters of this dissertation are based on experimental studies, each paper addresses a distinct research question and demonstrates a different use-case of experimental economic research [Roth, 1988; Charness, 2010]. Chapter 2 descriptively explores how individuals behave in a given decision context – reporting toward human versus machine auditors – without formulating directional hypothesis in advance (*"Establish Phenomena"*). Chapter 3 introduces a behavioral intervention– expert involvement in AI training – and *tests* its impact on advice adherence (*"Test Intervention"*). Chapter 4 seeks to transfer established concepts and phenomena from behavioral economics – overestimation and the hot hand fallacy – to an exemplary application case in another research field, being control engineering (*"Conceptual Transfer"*). Finally, Chapter 5 focuses on replicating existing experimental results – underplacement in relative self-assessment – and testing their robustness against a modified mechanism in the experimental design (*"Replication & Modification"*)).

Figure 1 visualizes the methodological and thematic intersections across the four research articles presented in the subsequent chapters, which are summarized in the following.

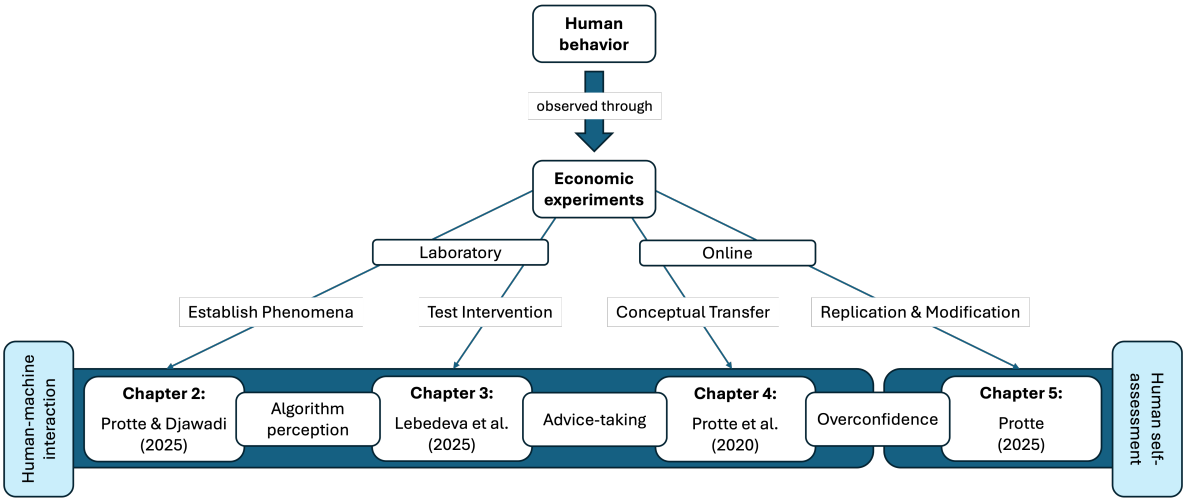


Figure 1: Overview of Dissertation Chapters' Methodological Approaches and Contents

Chapter 2: The first paper, **Protte & Djawadi (2025)**, investigates behavioral differences in dishonest conduct when individuals are confronted with either a human or a machine counterpart. While dishonesty is widely regarded as a trait inherent to human nature [Gerlach et al., 2019], and humans are well-known to cheat for personal gain or even on behalf of others, the prevalence of dishonest behavior varies sharply depending on the context. Between the extremes of complete honesty and outright dishonesty, many individuals engage in partial dishonesty, often by stretching or modifying the truth. This allows them to obtain additional benefits while maintaining a positive self-image or a perception of honesty [Jacobsen et al., 2018; Fischbacher and Föllmi-Heusi, 2013; Abeler et al., 2014, 2019; Peer et al., 2014; Bicchieri and Xiao, 2009; Khalmetski and Sliwka, 2019; Gneezy et al., 2015]. Although the topics of lying and cheating have been extensively studied in experimental psychology and behavioral economics, most research designs lack the real-world element of potential sanctions. This gap has recently been highlighted by a collaborative effort of leading researchers in the field [Shalvi et al., 2025]. They further emphasize the growing importance of accounting for emerging technological advancements – particularly AI and algorithmic systems – which have previously been identified as both enablers and mediators of dishonest human behavior [Köbis et al., 2021; Leib et al., 2024]. In response, they call for research to study how AI-based detection tools influence ethical behavior of individuals.

Our paper contributes to both of these research gaps by examining individual cheating behavior in human–machine and human–human interactions, under conditions where untruthful statements can be detected and penalized. To this end, we modify the classic die-roll paradigm by Fischbacher and Föllmi-Heusi [2013], introducing a lottery-based verification part that includes the threat of monetary punishment for dishonest reporting. The verification is conducted either by a human or an algorithm, and operates under either transparent or ambiguous audit rules. While we find the decision to engage in cheating to be primarily influenced by individual characteristics – namely gender and risk preferences – we observe a nuanced effect of verification entity on the extent of dishonest behavior. Under transparent verification rules, there is virtually no difference in cheating levels toward human and machine auditors. However, under opaque conditions, participants cheat significantly more when monitored by a machine entity, suggesting that this entity effect is largely driven by the algorithm's perceived "black-box" nature [Burrell, 2016; Tschider, 2020; Yeomans et al., 2019; Mahmud et al., 2022]. Consistent with prior literature, these findings may be explained by social image concerns: individuals may feel more compelled to behave ethically in the presence of a human auditor, a factor that is diminished or absent when interacting with a machine entity [Cohn et al., 2022; Biener and Waeber, 2024].

Our findings show that algorithmic systems may not discourage, but rather encourage dishonest behavior, particularly when their inner workings are non-transparent. They carry practical implica-

tions for both financial and non-financial audit procedures, as well as broader compliance, monitoring, and verification processes that hold potential for both automation and dishonest human behavior. The study underscores the need for careful design of AI-based oversight systems, emphasizing that procedural transparency is crucial for maintaining behavioral integrity and trust. We advise practitioners to exercise caution when fully automating verification in contexts where rule interpretation is complex or ambiguous, and to consider hybrid approaches – such as human-in-the-loop systems – particularly in settings where the economic cost of dishonesty exceeds the efficiency gains from automation.

Chapter 3: After Protte & Djawadi (2025) offer a descriptive analysis of how individuals interact with algorithmic entities in a specific economic setting in Chapter 2, the second paper, **Lebedeva et al. (2024)**, examines a behavioral intervention designed to influence decision-making in the context of algorithmic recommendation systems. A key methodological strength of economic experiments lies in their ability to assess the effects of organizational or policy interventions prior to real-world implementation, while such trials outside the laboratory would usually entail substantial financial costs or risks.

Automated Machine Learning (AutoML) presents a promising approach that enables end-users without prior data science knowledge to participate in the creation of machine learning (ML) models for algorithmic decision support systems [He et al., 2021; Singh and Joshi, 2022; Karmaker et al., 2020]. Major technology companies such as ("SageMaker Autopilot"), Microsoft ("Azure Machine Learning") or Google ("Vertex AI") have already introduced AutoML applications and products. However, based on results from behavioral research, AutoML represents a double-edged sword. While involving users in model creation may increase acceptance and adherence to algorithmic advice through enhancing perceived transparency and psychological ownership [Dietvorst et al., 2018; He and King, 2008; Jussupow et al., 2020; Norton et al., 2012; Sarstedt et al., 2016; Kawaguchi, 2021], greater expertise has been shown to foster algorithm aversion [Logg et al., 2019; Jussupow et al., 2020; Arkes et al., 1986]. We therefore investigate whether the positive effect of user involvement can offset the negative influence of user expertise. Specifically, we assess whether perceived control and understanding gained through participation in AI model training lead to increased adherence to algorithmic recommendations in an industrial decision-making context.

To test this, we conduct a laboratory experiment simulating a predictive maintenance scenario in a manufacturing setting. Participants assume the role of domain experts (engineers), making maintenance decisions based on machine indicators and non-binding AI recommendations. Productivity outcomes translate into monetary payoffs, while malfunctions – that can occur if maintenance is skipped – nullify productivity for that period. Maintenance ensures a reduced but certain payoff, creating a

realistic (though strongly simplified) trade-off. AI adherence is measured as the frequency with which participants followed AI advice that contradict their own initial judgments.

Our findings indicate that higher levels of involvement – ranging from no involvement, to passive observation of the training process, to active contribution of input – are associated with increased perceived influence over the AutoML system and greater self-reported understanding of its functionality. However, these perceptual differences do not translate into behavioral differences, as actual adherence to AI advice does not vary significantly across treatment groups. These results reflect prior findings gaps between stated trust in and actual reliance on algorithmic systems [Schmitt et al., 2021; Scharowski et al., 2023], as well as differences in the types of user involvement (on algorithm input versus algorithm output) [Gubaydullina et al., 2022]. A potential explanation consists in participants internalizing the model's performance as a reflection of their own input, reinforcing self-confidence and feeling of expertise, thus overriding the intended effect of the involvement manipulation. This is consistent with self-reports indicating high confidence and a strong sense of expertise among participants.

Overall, AI advice adherence exhibited considerable heterogeneity. On average, participants followed advice contradicting their own judgment in approximately 50% of cases. Notably, a substantial fraction of participants either consistently adhered to or entirely disregarded AI recommendations, indicating strong algorithmic appreciation or aversion, respectively.

In conclusion, while expert involvement in AutoML training improves attitudes toward algorithmic systems, it does not automatically translate into behavioral changes. These findings highlight the limitations of perception-based trust-building strategies, particularly when targeting expert users. For researchers, this challenges assumptions about the behavioral impact of perceived algorithmic transparency and influenceability. For practitioners, while user involvement may improve satisfaction and understanding, additional strategies – such as performance feedback or uncertainty framing – may be needed to influence adherence. Future research should explore methods to recalibrate expert confidence or enhance the perceived utility of AI recommendations in applied settings.

Chapter 4 Similar to Lebedeva et al.(2024) in Chapter 3, the third paper, **Protte et al. (2020)**, investigates behavioral responses to information provided by technological systems – this time in the context of feedback policy design. In doing so, it applies well-established insights from behavioral economics to the field of control engineering, where such perspectives have largely been overlooked in modeling human agents within control loops and cyber-physical systems.

Unlike traditional control models that treat human agents as Markov decision-makers [Puterman, 1994; van Otterlo and Wiering, 2012] and thus implicitly rational or deterministic, this paper acknowledges bounded rationality [Simon, 1972; Selten, 2001] and seeks to capture perceptual and behavioral

limitations, such as overconfidence [Svenson, 1981; Moore and Healy, 2008; Healy and Moore, 2007] and the hot hand fallacy [Gilovich et al., 1985; Camerer, 1989; Croson and Sundali, 2005], that influence real-world decisions. We introduce an experimental framework to empirically study these behavioral effects within control systems. The experiment simulates a human-in-the-loop environment via a surveillance drone piloting task inspired by Feng et al. [2016]. Participants must decide how many times a drone should fly over ten traffic junctions to maximize information value, while each flight carries a small but compounding risk of the drone crashing – an event that would eliminate the information collected during that period and nullify all subsequent payoffs. Two feedback conditions are compared: a closed feedback loop (providing real-time image quality feedback after each flight) and an open loop (requiring all flight decisions to be made in advance without interim feedback).

Results show that participants deviate from optimal strategies predicted by classical control theory [Bertsekas, 2005; van Overloop et al., 2015; Ercan et al., 2017; Inoue and Gupta, 2019]. The data reveals a bimodal distribution of behavior: some participants acted overconfidently by flying excessively, while others were overly conservative. Across both treatments, subjects frequently flew more rounds than optimal, increasing crash risk. Immediate feedback under the closed-loop condition improved average performance (i.e., greater information value accumulated), yet also exacerbated overconfidence and susceptibility to a "hot hand". Notably, despite the majority of subjects being classified as risk-averse [per Dohmen et al., 2011], over 80% continued flying beyond the optimal point after receiving positive sequential feedback. This behavior is consistent with the hot hand fallacy, suggesting that subjects misinterpreted random positive outcomes as indicative of skill or predictability. Conversely, the absence of interim feedback in the open-loop condition led to underconfident tendencies in many participants.

The study thus provides evidence that real-time feedback can paradoxically amplify behavioral biases rather than mitigate them. Overall, the paper makes a strong case for incorporating behavioral economic principles into the design of human-in-the-loop systems. It cautions against relying solely on normative decision models and highlights the importance of treating human decision-makers as dynamic and imperfectly rational components in control settings. Scientifically, the study challenges the assumptions underpinning Markov Decision Processes and Bayesian rationality in human-centered control models. It advocates for behaviorally-informed frameworks that better reflect human behavior. Practically, the findings underline the critical importance of intelligent feedback design. Blindly increasing feedback frequency or granularity may inadvertently worsen outcomes due to cognitive biases. System designers should explore intermittent or tailored feedback schemes to mitigate these adverse effects.

Similar to the scenario described in Chapter 3, testing alternative feedback policies in real-world settings would be highly costly due to the potential for disruption during live system operation, if

not completely infeasible given the risk of drone crashes. This paper helps bridge the gap between behavioral research and technological application by providing a controlled experimental framework. It serves as an introductory reference for interdisciplinary research on human behavior in cyber-physical systems, offering valuable insights for those seeking a deeper understanding of the human component in human-in-the-loop modeling.

Chapter 5: While Protte et al. (2020) in Chapter 4 adopt a more traditional and simplified approach to overconfidence in the form of overestimation – operationalizing it as the discrepancy between an objectively optimal decision strategy and the one actually implemented, primarily to facilitate its conceptual transfer to control engineering – the fourth paper, **Protte (2025)**, delves into the phenomenon with greater depth. Focusing on overplacement, a specific form of overconfidence characterized by inflated self-assessments relative to others [Larrick et al., 2007; Moore and Healy, 2008],´the study replicates and adapts a seminal experiment by Erik Hoelzl and Aldo Rustichini [2005]. Despite their importance being consistently emphasized, replications of economic experiments are rarely actually conducted (and published) [Charness, 2010; Rosenthal, 1990; Weimann and Brosig-Koch, 2019]. The original study reported an unexpected pattern of underplacement in difficult tasks under real monetary incentives, contradicting the prevailing notion of a general human tendency toward overconfidence [DeBondt and Thaler, 1995]. The principal aim of the replication is to test the robustness of those findings by introducing a subtle yet methodologically important treatment modification.

In their design, Hoelzl and Rustichini use a voting decision between two bonus payment mechanisms – one performance-based (rewarding the top 50% of participants) and one lottery-based (offering a 50% chance of winning via a die roll) – as a proxy for self-assessed relative performance. However, this study questions whether these two schemes are truly comparable. While the performance-based mechanism involves interdependent success probabilities and a number of winners that is fixed ex-ante, the lottery mechanism features independent success probabilities and allows for a variable number of winners. These structural differences may lead participants to perceive the two options differently, based not only on expected monetary outcomes but also on non-monetary preferences or motivations, such as aversion against ambiguity, risk or inequality, altruism and fairness considerations [Benoît et al., 2022; Gneezy et al., 2015; Strang and Schaube, 2025; Andreoni and Miller, 2002; Loewenstein et al., 1989]. As a result, the extent of inaccurate self-assessment inferred from voting behavior may be overstated due to design artifacts in the elicitation method. Therefore, this replication study aims to determine whether the observed underplacement effect stems from biased self-assessments, from features of the experimental design (the lottery mechanism), or from other, yet unidentified factors. To do so, it compares the original lottery format with an alternative design that matches the structural properties

of the performance-based scheme, by randomly awarding bonuses to exactly 50% of participants. Additionally, it includes a self-report component to elicit participants' underlying voting motives and help disentangle potential explanations for the observed results.

Contrary to the original study, participants in both treatment conditions exhibit clear signs of overplacement, with nearly three-fourths of subjects preferring the performance-based payoff. However, the change in lottery design does not appear to affect voting outcomes. Generally, voting behavior is strongly determined by participants' expectations of their relative performance – key predictors include both individual and group performance estimates. Nevertheless, non-monetary motives and preferences also play a decisive role. In addition to actual self-assessed confidence, three primary decision rationales emerge: (1) normative beliefs, such as principle of merit among performance voters and equal opportunity among lottery voters, (2) a preference for control (among performance voters only), and (3) performance feedback from sample questions, despite the fact that it does not provide information on relative performance. Furthermore, gender and educational status are found to significantly influence voting behavior, while attitudes toward risk and ambiguity, social comparison tendencies, self-efficacy, and altruism show no meaningful association with voting behavior. These findings contribute to the methodological discourse on measuring overconfidence and provide insights into alternative drivers of discrepancies from payoff-maximizing calibration in self-assessments. In doing so, the study challenges the robustness of Hoelzl and Rustichini's original findings regarding underplacement in vote-based elicitation mechanisms involving difficult tasks under monetary incentives.

This replication demonstrates that methodological choices – particularly in the design of elicitation mechanisms such as votes, tests, and lotteries – can substantially shape the interpretation of overconfidence. It supports prior methodological critiques suggesting that overplacement measures based on choice behavior may reflect underlying preferences for control [Owens et al., 2014; Benoît et al., 2022] or fairness [Krawczyk, 2012; Strang and Schaube, 2025], rather than – or in addition to – genuine miscalibration. Ultimately, the study enhances the internal validity of behavioral confidence measures and helps distinguish between "apparent" and "true" overconfidence [Benoît and Dubra, 2011; Benoît et al., 2014]. These insights have practical implications for the design of compensation schemes and personnel assessment tools. Recognizing that self-assessments in high-stakes environments may reflect strategic or normative considerations rather than pure confidence can inform the development of more effective and fair incentive structures.

Table 1 on the following page provides a concise overview of the studies (chapters) in this dissertation, briefly summarizing their topics addressed, research objectives, sampling, key findings, and main contributions. The breakdown of individual co-author contributions for each study, as well as publication status and presentations given (as of July 2025), follows in Section 1.5.

Table 1: Chapter Overview

Research Objective	Findings & Contributions	Topics	Method & Sample
Chapter 2: Human vs. Algorithmic Auditors: The Impact of Entity Type and Ambiguity on Human Dishonesty Marius Protte & Behnud Mir Djawadi (2025)			
Do individuals cheat more when they are audited by a machine or another human? How does ambiguity about verification rules ("black-box") affect dishonesty toward audits conducted by a machine?	• Under transparent rules, levels of dishonesty are practically equal (in frequency and magnitude) toward human and machine auditors • Dishonesty toward both entities becomes more polar under opaque audit rules (ambiguity), partial cheating is nearly eliminated • Under opaque audit rules, the magnitude of cheating is significantly higher with a machine auditor, likely due to social image considerations • Trade-off in moral domains: Potential adverse effects through more unethical behavior vs. efficiency improvements through automation	Cheating behavior; Opacity; Algorithm aversion; Algorithm Appreciation; Social Image	• Incentivized laboratory experiment (BaER-Lab) • Student sample, 4 treatments ($N = 170$)
Chapter 3: Does the involvement of domain experts in the AI training affect their AI perception and AI adherence? An experiment in the context of industrial AutoML applications Anastasia Lebedeva, Marius Protte, Dirk van Straaten & René Fahr (2024)			
Can domain experts' adherence to algorithmic advice be improved through their involvement during model training – typically competing effects of user involvement (+) and expertise (−) on adherence (e.g., Jussupow et al. 2020) – given that involvement does not change model accuracy?	• No treatment differences in algorithm adherence are observed between active, passive and non-involvement • Heterogeneous adherence patterns overall, most participants either always follow advice or never do • Experts likely internalize training success • Ready-made products may be equally or more efficient than co-created models from a pure cost-benefit perspective	Algorithm aversion; Experts; Involvement; Automated Machine Learning	• Incentivized laboratory experiment (BaER-Lab) • Student sample, 3 treatments ($N = 154$)
Chapter 4: Behavioral Economics for Human-in-the-Loop Control Systems Design: Overconfidence and the Hot Hand Fallacy Marius Protte, René Fahr & Daniel E. Quevedo (2020)			
How do behavioral patterns like overconfidence and the hot hands affect optimization processes in control engineering? How do these behavioral patterns interact with feedback policies in human-in-the-loop control systems?	• Behavioral inefficiencies are illustrated in a surveillance drone piloting scenario, as established insights from behavioral economics have been largely neglected in control engineering optimization problems • Overestimation and hot hand patterns (erroneous extrapolation of random sequence) are observed in closed feedback loops that would pose substantial risk of economic damage in real-world application • Inefficient conservatism is observed for open feedback loops	Humans in cyber-physical systems; Feedback policy design; Closed loop; Open loop; Behavioral biases	• Incentivized online experiment (BaER-Lab pool) • Student sample, 2 treatments ($N = 105$)
Chapter 5: Explaining Apparently Inaccurate Self-Assessments of Relative Performance: A replication and adaptation of "Overconfident: Do you put your money on it?" by Hoelzl & Rustichini (2005) Marius Protte (2025)			
Do lottery design characteristics lead to the observation of (apparent) underconfidence in the original experiment by Hoelzl & Rustichini (2005)? Which factors besides confidence biases explain relative misplacement in performance self-assessments?	• H&R's findings of underplacement are not replicated • Applying an alternative lottery mechanism does not alter results • Self- and group-assessments are strong and significant predictors of payoff-mechanism preference; majority appears accurately calibrated • Signals through sample questions, preferences for control, and normative beliefs (merits & equal chances) are identified as main decision drivers besides confidence	Overconfidence; Overplacement; Better-than-average-effect; Lottery Design; Control preference; Fairness	• Incentivized online experiment (Prolific) • Mixed sample, 2 treatments ($N = 200$) + 1 robustness check ($n = 100$)

1.5 Contributions to Joint Work with Co-authors

Chapter 2	**Protte, M. and Djawadi, B. M. (2025):** **Human vs. Algorithmic Auditors: The Impact of Entity Type and Ambiguity on Human Dishonesty** **Frontiers in Behavioral Economics, 4, 1645749**
Contribution of joint work with co-authors	• Co-authorship with Dr. Behnud Mir Djawadi (65% M. Protte, 35% B. M. Djawadi) • Idea and experimental design were jointly developed • Programming of experiment and data collection by M. Protte • Data analysis by M. Protte • Write-up of paper jointly
Conferences / Workshops	• Faculty Research Workshop 2022, Melle, Germany (poster presentation) • Annual Meeting 2023 of "Gesellschaft für experimentelle Wirtschaftsforschung e.V." (GfeW), Erfurt, Germany • International Association for Research in Economic Psychology/Society for the Advancement of Behavioral Economics (IAREP-SABE) joint Conference 2024, Dundee, Scotland • LEE Workshop "Human-AI interaction from an economics perspective", Benicasim, Spain
Scientific dissemination	• Start of work: July 2022 • Experiment pre-test: September 2022 • Data collection in June 2023 and October 2023 • First Draft: June 2024 • Published in *Frontiers in Behavioral Economics* in August 2025

Chapter 3	**Lebedeva, A., Protte, M., van Straaten, D. and Fahr, R. (2024): Does the involvement of domain experts in the AI training affect their AI perception and AI adherence? An experiment in the context of industrial AutoML applications Advances in Information and Communication: Proceedings of the 2024 Future of Information and Communication Conference (FICC), 178-204. Springer Nature**
Contribution of joint work with co-authors	• Co-authorship with Dr. Anastasia Lebedeva, Dr. Dirk van Straaten, and Prof. Dr. René Fahr (35% A. Lebedeva, 35% M. Protte, 15% D. v. Straaten, 15% R. Fahr) • Idea and experimental design were jointly developed • Programming of experiment and data collection by A. Lebedeva and M. Protte • Data analysis by A. Lebedeva and M. Protte • Write-up by A. Lebedeva and M. Protte • Comments and corrections by D. v. Straaten and R. Fahr
Conferences / Workshops	• International Association for Research in Economic Psychology/Society for the Advancement of Behavioral Economics (IAREP-SABE) joint Conference 2023, Nice, France • Faculty Research Workshop 2023, Paderborn, Germany (poster presentation) • Future of Information and Communication Conference (FICC) 2024, Berlin, Germany • Annual Meeting 2024 of "Gesellschaft für experimentelle Wirtschaftsforschung e.V." (GfeW), Cologne, Germany
Scientific dissemination	• Start of work: May 2022 • Data collection in December 2022 and April 2023 • First Draft: June 2023 • Current Draft: Published in *Advances in Information and Communication* in April 2024 • **Note**: This research project was also part of the dissertation of my co-author Anastasia Lebedeva.

Chapter 4	**Protte, M., Fahr, R. and Quevedo, D. E. (2020):** **Behavioral Economics for Human-in-the-Loop Control Systems Design: Overconfidence and the Hot Hand Fallacy** **IEEE Control Systems Magazine, 40(6), 57-76**
Contribution of joint work with co-authors	• Co-authorship with Prof. Dr. René Fahr and Prof. Dr. Daniel E. Quevedo (40% M. Protte, 30% R. Fahr, 30% D. E. Quevedo) • Idea and experimental design were jointly developed • Programming of experiment and data collection by M. Protte • Data analysis M. Protte and R. Fahr • Write-up by M. Protte and D. E. Quevedo
Conferences / Workshops	• Annual Meeting 2019 of "Gesellschaft für experimentelle Wirtschaftsforschung e.V." (GfeW), Düsseldorf, Germany (presentation by R. Fahr)
Scientific dissemination	• Start of work: April 2019 • First draft: February 2020 • Published in *IEEE Control Systems Magazine* in December 2020

Chapter 5	**Protte, M. (2025):** **Explaining Apparently Inaccurate Self-Assessments of Relative Performance: A Replication and Adaptation of "Overconfident: Do you put your money on it?" by Hoelzl & Rustichini (2005)** **Working Paper, available at: https://doi.org/10.48550/arXiv.2507.15568; submitted to Journal of Economic Psychology**
Contribution of joint work with co-authors	This work is single-authored.
Conferences / Workshops	• Society for the Advancement of Behavioral Economics (SABE) Conference 2025, Trento, Italy • TIBER Symposium on Psychology and Economics 2025, Tilburg, Netherlands
Scientific dissemination	• Start of work: January 2024 • First draft: February 2025 • Current Draft: Submitted to *Journal of Economic Psychology* in 2025

Chapter 2:

"Human vs. Algorithmic Auditors: The Impact of Entity Type and Ambiguity on Human Dishonesty"

Human vs. Algorithmic Auditors: The Impact of Entity Type and Ambiguity on Human Dishonesty

Marius Protte[*,†], Behnud Mir Djawadi[*]

Abstract

Human-machine interactions become increasingly pervasive in daily life and professional contexts, motivating research to examine how human behavior changes when individuals interact with machines rather than other humans. While most of the existing literature focused on human-machine interactions with algorithmic systems in advisory roles, research on human behavior in monitoring or verification processes that are conducted by automated systems remains largely absent. This is surprising given the growing implementation of algorithmic systems in institutions, particularly in tax enforcement and financial regulation, to help monitor and detect misreports, or in online labor platforms widely implementing algorithmic control to ensure that workers deliver high service quality. Our study examines how human dishonesty changes when detection of untrue statements is performed by machines versus humans, and how ambiguity in the verification process influences dishonest behavior. We design an incentivized laboratory experiment using a modified die-roll paradigm where participants privately observe a random draw and report the result, with higher reported numbers yielding greater monetary rewards. A probabilistic verification process introduces risk of detection and punishment, with treatments varying by verification entity (human vs. machine) and degree of ambiguity in the verification process (transparent vs. ambiguous). Our results show that under transparent verification rules, cheating magnitude does not significantly differ between human and machine auditors. However, under ambiguous conditions, cheating magnitude is significantly higher when machines verify participants' reports, reducing the prevalence of partial cheating while leading to behavioral polarization manifested as either complete honesty or maximal overreporting. The same applies when comparing reports to a machine entity under ambiguous and transparent verification rules. These findings emphasize the behavioral implications of algorithmic opacity in verification contexts. While machines can serve as effective and cost-efficient auditors under transparent conditions, their black box nature combined with ambiguous verification processes may unintentionally incentivize more severe dishonesty. These insights have practical implications for designing automated oversight systems in tax audits, compliance, and workplace monitoring.

JEL Classification: C91, D81, D91, M42

Keywords: dishonesty; cheating; ambiguity; human-machine interaction; algorithm aversion; algorithm appreciation

---

[*]Paderborn University, Heinz-Nixdorf-Institute, Fürstenallee 11, 33102 Paderborn

[†]Corresponding author, marius.protte@upb.de

An updated version of this article has been published in *Frontiers in Behavioral Economics* after the submission of this dissertation, DOI: https://doi.org/10.3389/frbhe.2025.1645749. This research was funded by the Deutsche Forschungsgemeinschaft within the "SFB 901: On-The-Fly (OTF) Computing – Individualised IT-Services in Dynamic Markets" program (160364472).

2.1 Introduction

Human-machine interaction is ubiquitous in today's world, driven by increasing automation and the growing reliance on algorithms and artificial intelligence (AI) in decision-making. AI, algorithmic advisors, and computerized decision support systems are employed in various domains, where they often outperform human judgment. Notable examples include medicine and healthcare [Cheng et al., 2016; Gruber, 2019], public administration [Kouziokasa, 2017; Bignami, 2022], autonomous driving [Levinson et al., 2008], human resource management [Highhouse, 2008], investment decisions [Tao et al., 2021], insurance claim processing [Komperla, 2021], tax audits [Black et al., 2022; Baghdasaryan et al., 2022], and criminal jurisdiction [Kleinberg et al., 2018], among others. At the same time, demographic shifts and skilled labor shortages present pressing societal challenges, which are increasingly addressed through algorithmic and AI-based automation.

Despite algorithms often demonstrating superior predictive accuracy compared to human forecasters, people frequently prefer human input when given a choice between algorithmic and human forecasts [Dietvorst et al., 2015]. Likewise, individuals regularly disregard algorithmic advice in favor of their own judgment, even when doing so is not rational and leads to inferior outcomes [Burton et al., 2019; Jussupow et al., 2020]. Conversely, the perceived reliability, consistency, and objectivity of algorithms can lead to over-reliance on their advice, particularly in structured and predictable tasks [Klingbeil et al., 2024; Banker and Khetani, 2019]. This duality in perception highlights the complexity of human attitudes toward machine-supported decision-making, as levels of algorithm acceptance and adherence typically vary widely across individuals and contexts [Fenneman et al., 2021].

Many of the fields of application mentioned at the beginning inherently involve moral considerations to which individual differences in the perception of humans versus machines pertain. When algorithms act as ethical advisors, an asymmetry in their impact becomes apparent: algorithmic advice appears largely unsuccessful in promoting honest behavior, but is able to facilitate dishonest behavior [Leib et al., 2024]. Similarly, AI agents can function as enablers of unethical behavior in decisions that can be delegated by offering individuals a means to outsource or share the moral load imposed by unethical behavior [Köbis et al., 2021; Bartling and Fischbacher, 2012]. Regarding honesty, Cohn et al. [2022] find significantly more cheating when individuals interact with machines than with humans, regardless of whether the machine has anthropomorphic features. Dishonest individuals actively prefer machine interaction when given an opportunity to cheat. Meanwhile, people cheat less in the presence of a robot [Petisca et al., 2022] or digital avatar [Mol et al., 2020] if it signals awareness of the situation than when being alone, even when it cannot intervene.

However, what happens to human dishonest behavior if machines can detect when someone lies

or makes an untrue statement? Does behavior potentially change because of the machine entity itself or because of the ambiguity machines create through their "black box" nature? Concurrent with the tendency to use AI as advisors, algorithms are also used to monitor human conduct. For example, there is growing implementation of algorithmic systems in institutions, particularly in tax enforcement and financial regulation, to help monitor and detect misreports [e.g., Faúndez-Ugalde et al., 2020]. Similarly, online labor platforms widely implement algorithmic control to ensure that workers consistently deliver high quality services [Wang et al., 2024]. Despite the prevalence and impact of this form of human-machine interaction, we have limited understanding of how human dishonest behavior is shaped when their actions are subject to machine verification. We therefore ask the following research questions:

How does human dishonesty change when detection of untrue statements is performed by machines versus humans, and to what extent does ambiguity in the verification process influence dishonest behavior?

We hereby make two important contributions. First, our research extends findings from the dishonesty literature by investigating scenarios where machines serve not as advisors or partners but as verification entities that detect untrue statements, an increasingly common human-machine interaction context. Second, while institutions such as tax authorities have increasingly implemented algorithmic systems to identify suspicious patterns in tax reports, our research clarifies whether the use of such machines creates a deterrence effect that reduces dishonesty. These insights may also provide valuable information for organizations implementing monitoring systems, where research regularly shows that electronic surveillance systems are often perceived negatively by employees and can even be associated with increased employee intentions to engage in counterproductive workplace behaviors.

To answer our research questions, we conduct an incentivized one-shot laboratory experiment that employs a modified version of the die-roll paradigm introduced by Fischbacher and Föllmi-Heusi [2013]. Participants privately observe a random draw and report its outcome, with monetary payoffs tied to the reported number – creating an opportunity to profit from dishonesty. We introduce a two-stage verification process in which reports that may turn out to not coincide with the truth are sanctioned with a substantial monetary penalty. By incorporating elements of risk and uncertainty into the traditional dishonesty paradigm, our methodological approach maintains a generalizable framework that intentionally abstracts from domain-specific settings such as tax evasion or corruption. While these contexts share similar mechanisms of detecting and sanctioning deviant behavior, they frequently involve additional motivational factors such as civic duty, moral obligations, and imposing negative externalities on others that could confound the fundamental relationship between dishonest behavior and verification entity that we aim to isolate. We vary both the verification entity (Human vs. Machine)

and the level of ambiguity involved in processing the die-roll reports (Black box vs. Transparent) to compare how participants' dishonest behavior is affected by who verifies their reports and how transparent the verification process is. We control for factors such as risk preferences, attitudes toward ethical dilemmas, perceived closeness to the auditor, and technology affinity.

The proceeding paper is structured as follows: Section 2.2 reviews prior research on perceptions of algorithmic entities, human dishonesty, and their intersection. With this context established, two hypotheses are derived for the experimental study. Subsequently, Section 2.3 outlines the experimental design and procedure in detail. Section 2.4 presents descriptive results, followed by hypothesis testing and multivariate regression analysis. Finally, Section 2.5 offers an interpretation of the findings and concludes with a discussion of the study's limitations and implications.

2.2 Related Literature and Derivation of Hypotheses

2.2.1 Literature Overview

Algorithm Perception Recent advances in human-machine interaction research increasingly focus on how individuals perceive algorithms and AI, particularly in the context of algorithm aversion and algorithm appreciation [e.g., Mahmud et al., 2022; Jussupow et al., 2020; Dietvorst et al., 2015, 2018; Castelo et al., 2019; Logg et al., 2019; Fuchs et al., 2016][1]. Within this literature, the term *algorithm* is often used as a broad synonym, encompassing various technological systems, including decision support systems, automated advisors, robo-advisors, digital agents, machine agents, forecasting tools, chatbots, expert systems, and AI-generated decisions [Mahmud et al., 2022][2]. In line with this, we use the term "algorithm" to denote any technological system that applies a deterministic, stepwise process to decision-making [Dietvorst and Bharti, 2020].

Generally, attitudes toward algorithms vary widely among individuals. These attitudes are not fixed, but rather context-dependent, reflecting both algorithm aversion and algorithm appreciation [Fenneman et al., 2021; Hou and Jung, 2021]. *Algorithm aversion* describes the tendency – whether conscious or unconscious – to resist relying on algorithms, even when they are demonstrably outperform human judgment. People frequently reject algorithmic advice in favor of their own or other humans' opinions, despite being aware of the algorithm's superior accuracy and incurring material costs for doing so [Dietvorst et al., 2015, 2018; Mahmud et al., 2022; Jussupow et al., 2020]. Although people

[1] Empirical research in this field can be broadly categorized into two strands: (1) studies in which humans interact with algorithms, programs, chatbots, or AI systems through a computer interface [e.g., Cohn et al., 2022; Biener and Waeber, 2024; Dietvorst et al., 2015; Logg et al., 2019]; and (2) studies involving humans interacting with anthropomorphic robots, focusing on perceived trustworthiness, intelligence, or reciprocity – often observed from a third-person perspective [e.g., Canning et al., 2014; Ullman et al., 2014; Sandoval et al., 2020]. The present study is concerned solely with the former type of interaction.

[2] From a technical standpoint, an algorithm is defined as a sequential logical process applied to a data set to accomplish a certain outcome. This process is automated and processes without human interference [Gillespie, 2016].

frequently attribute near-perfect performance to algorithms [Dzindolet et al., 2002], they are quicker to lose trust in them following errors, regardless of the error's context or severity [Renier et al., 2021]. In contrast, equivalent human mistakes are more readily excused [Madhavan and Wiegmann, 2007]. Conversely, *algorithm appreciation* refers to situations in which individuals are more likely to follow identical advice when it originates from an algorithm rather than a human, often displaying greater confidence in such recommendations despite having little to no insight into the algorithm's internal workings [Logg et al., 2019]. This effect is especially pronounced when the algorithm signals expertise [Hou and Jung, 2021]. A systematic literature review by Mahmud et al. [2022] concludes that algorithm acceptance varies along several demographic lines: older individuals and women tend to show greater aversion, while higher education is associated with greater acceptance. Moreover, algorithm aversion is often more pronounced among domain experts [Logg et al., 2019; Jussupow et al., 2020].

Both these directions of biased algorithm perception may result in economic inefficiencies. On the one hand, algorithms, despite not being entirely free of errors, consistently provide more accurate decisions than human counterparts [Dawes et al., 1989; Logg et al., 2019]. Yet, in decisions under risk and uncertainty, individuals often disregard even high-quality algorithmic advice due to heightened sensitivity to potential errors, leading to suboptimal outcomes [Dietvorst and Bharti, 2020; Prahl and Swol, 2017; Jussupow et al., 2020]. This reluctance is particularly evident in morally salient domains – such as medicine, criminal justice, or military contexts – where algorithmic input is frequently rejected even when it aligns with human decisions and produces efficient outcomes [Bigman and Gray, 2018]. On the other hand, unreflective algorithm appreciation may results in over-reliance, where individuals defer to algorithmic recommendations despite contradictory contextual knowledge or better judgment. This can lead to suboptimal decisions with unintended consequences for both the decision-maker and affected third parties [Klingbeil et al., 2024]. For example, Banker and Khetani [2019] find that consumers often rely to heavily on algorithmic recommendations, leading to inferior purchasing decisions. Similarly, Krügel et al. [2022] demonstrate that individuals' decision-making in ethical dilemmas can be manipulated through overtrust in AI. Two key factors determining an individual's unique degree of algorithm adherence, i.e., their inclination to either use or avoid algorithms, are anticipated efficacy and trust placed in the algorithmic system [Fenneman et al., 2021]. Perceived efficacy appears to have a stronger positive influence on willingness to rely on algorithms than discomfort or unease associated with using them [Castelo et al., 2019]. In terms of trust, similar factors as in human relationships – perceived competence, benevolence, comprehensibility, and responsiveness – also apply to automation. Additionally, perceptions specific to technology, such as reliability, validity, utility, and robustness, play an important role [Hoffman et al., 2013].

Human Dishonesty People lie and cheat for their own benefit or for the benefit of others [Abeler et al., 2019; Jacobsen et al., 2018]. However, despite being able to maximize their monetary payoffs, people often abstain from lying and cheating, for various reasons, e.g., general preferences for truth-telling, intrinsic lying costs, lying aversion, emotional discomfort and social image concerns [Abeler et al., 2014, 2019; Bicchieri and Xiao, 2009; Khalmetski and Sliwka, 2019]. Additionally, lying behavior differs in magnitude, distinguishing between full liars (i.e., lying to the maximum extent possible), partial liars (i.e., exaggerating the actual outcome but not to the maximum), and fully honest individuals [Fischbacher and Föllmi-Heusi, 2013; Gneezy et al., 2018]. Fittingly, previous experimental research (either in the lab or field) finds a considerable variance in cheating behavior among individuals with the opportunity to do so. Observed proportions of fully honest decision-making usually range between 40 [Fischbacher and Föllmi-Heusi, 2013] and close to 70 percent [Peer et al., 2014; Djawadi and Fahr, 2015; Gneezy et al., 2018], while [Abeler et al., 2014] observe close to no cheating at all. The large-scale meta-study by Gerlach et al. [2019] finds cheating rates of approximately 50% across common experimental lying and cheating settings (sender-receiver games, die-roll tasks, matrix tasks). Meanwhile, similar heterogeneity can be found for the respective degree of dishonesty, as fractions of 2.5% and 3.5% lying to the maximum extent possible are observed by Shalvi et al. [2011] and Peer et al. [2014] respectively, while around 20% of individuals lie to the maximum extent possible in Fischbacher and Föllmi-Heusi [2013]. Gneezy et al. [2018] find up to 47% of subjects lying, and up to 91% doing so to the maximum extent possible, depending on the combination of given reporting mechanism and having the opportunity to do so, as the degree of cheating generally appears to vary heavily with personal and situational factors [Gerlach et al., 2019].

The possibility of lying and cheating in (nearly) all domains of human-machine interaction mentioned above imposes ethical challenges and financial costs to both businesses and society. Cohn et al. [2022] find that individuals are more likely to engage in dishonest behavior when interacting with a machine rather than a human, regardless of whether the machine exhibits human-like characteristics. Moreover, individuals with an intention to cheat tend to prefer interacting with machines over humans. These patterns are largely attributed to diminished social image concerns and the perception that machines possess lower levels of agency [Cohn et al., 2022; Biener and Waeber, 2024].

However, these findings stem from situations where untrue statements cannot be detected. In daily and economic life, such perfect concealment cannot always be guaranteed, and the recipient of a false statement might discover the truth. As machines may be perceived as more accurate than humans at detecting untrue statements, their presence as verifiers could potentially reduce cheating compared to human verification. Thus, the findings of the existing dishonesty literature may not apply to situations where detection is possible, necessitating empirical investigation of this specific context.

2.2.2 Hypotheses

Referring to the literature on algorithm aversion and appreciation, it becomes evident that in numerous daily and economic contexts, functionally equivalent actions performed by humans and machines can be differently perceived by human recipients. For the examination of detecting and potentially sanctioning dishonest behavior, there also exist competing arguments regarding whether dishonesty rates might increase or not when machines rather than humans verify the statements' truthfulness. On the one hand, algorithmic decisions are usually being perceived as more objective, consistent and less error-prone [Dzindolet et al., 2002, 2003; Renier et al., 2021]. Human individuals intending to engage in dishonest behavior may therefore prefer human verification of their reports, anticipating a higher chance of avoiding detection and subsequent sanctions due to perceived limitations in human monitoring capabilities. Further, individuals may be more likely to act dishonestly when humans verify their statements because they believe humans exercise discretionary judgment based on empathy or fairness considerations. Such perceptions have been observed particularly in morally charged contexts [Dietvorst et al., 2015; Mahmud et al., 2022; Jauernig et al., 2022]. Machines, conversely, are conceptualized as rigid rule-followers lacking such affective capacities [Haslam, 2006; Bigman and Gray, 2018; Gogoll and Uhl, 2018; Niszczota and Kaszás, 2020]. On the other hand, empirical evidence indicates that human individuals perceive algorithmic surveillance more negatively than human surveillance [Schlund and Zitek, 2024]. Further, related literature provides indirect evidence that algorithmic monitoring does not prevent but in some cases even facilitate deviant behaviors. For instance, Wang et al. [2024] analyze data from a ride-hailing platform and finds that intensified algorithmic control implemented through work-related monitoring positively influences customer-directed deviant behavior among drivers. Similarly, Liu et al. [2021] compare conventional taxi and Uber drivers, finding that despite enhanced algorithmic tracking capabilities in the latter context, route manipulation through detours that benefits drivers at passengers' expense is more prevalent in Uber rides compared to taxi rides during surge pricing periods. More direct evidence comes from experimental economics. Cohn et al. [2022] find that individuals are significantly more likely to cheat machine agents than human ones, regardless of the medium (voice or text) or whether the machine features anthropomorphic traits. Dishonest individuals also show a preference for interacting with machines when given an opportunity to cheat. This behavior is attributed to social image concerns in interactions between humans, which have been previously identified as a key inhibitor of dishonesty [Abeler et al., 2019; Khalmetski and Sliwka, 2019]. Similar findings are reported by Biener and Waeber [2024], who observe greater honesty when participants report the outcomes of unobserved, payoff-relevant random draws to a human rather than a chatbot. The degree of perceived agency, as well as considerations of social image and norms, appear to drive this difference. Social image concerns represent a plausible factor in our setting as

well. Being detected and sanctioned by another human may carry higher reputational consequences for the individual than when detection occurs through algorithmic means, as machines are less likely to be perceived as forming judgments about character or moral worth. This asymmetry would suggest that algorithmic verification systems may inadvertently facilitate dishonest behavior by lowering the social costs that typically deter such conduct when human oversight is present. Given these competing arguments, we formulate our first hypothesis in a conservative manner without specifying the direction of potential behavioral differences:

Hypothesis 1: Human dishonest behavior will differ when their statements' truthfulness is verified by humans or machines.

As technological trends suggest that machines will increasingly be employed for automated detection processes, our second hypothesis focuses on machines as verification entities. Beyond psychological, biological, and ethical dimensions, perceptual differences between humans and machines are typically rooted in technological characteristics, where a central debate concerns whether algorithmic systems should operate through transparent rules or be deliberately kept ambiguous. There are indications that this discussion is also relevant for the human-machine interaction in our setting. As algorithms, by nature, tend to be opaque rather than transparent, they are frequently perceived as "black boxes" that convert some type of input into some type of output without revealing their internal logic [Tschider, 2020; Mahmud et al., 2022]. Commonly, humans neither understand nor are aware of how algorithms function, which constitutes a major reason for them rejecting algorithms and their advice [Yeomans et al., 2019; Dzindolet et al., 2002; Kayande et al., 2009; Mahmud et al., 2022]. From the perspective of advice-taking, "opening the black box" through increasing transparency, accessibility, explainability, interactivity and tunability has been widely advocated to foster trust in and reduce aversion toward algorithms [Sharan and Romano, 2020; Chander et al., 2018; Holzinger et al., 2017; Litterscheidt and Streich, 2020; Shin, 2020]. However, it has been shown that even if an algorithm's underlying logic is disclosed to the decision-maker, it may remain unintelligible, especially to non-experts [Önkal et al., 2009]. Decision context also plays a crucial role. Sutherland et al. [2016] find that humans are more inclined to rely on algorithms in uncertain environments. Contrastingly, Longoni et al. [2019] report greater aversion to algorithmic decision-making in high-stakes environments rife with uncertainty such as healthcare. These mixed findings reflect a distinction in how humans perceive decisions under ambiguity (i.e., uncertainty) differently from decisions under risk, where potential outcomes and related probabilities are known [Ellsberg, 1961; Einhorn and Hogarth, 1986; Fox and Tversky, 1995; Chow and Sarin, 2001]. The influence of an algorithm's black box nature specifically on human dishonest behavior is therefore not straightforward. Under transparent verification rules, dishonesty may reflect a rational

cost-benefit analysis based on known probabilities, on the basis of which partial cheating nay reflect a rational outcome. Under ambiguity, however, where the likelihood of being detected and punished is unknown, such estimates become difficult. Therefore, transparency might actually encourage more dishonesty compared to an ambiguous detection process, as individuals can better assess these risks. In contrast, when detection probability parameters are unavailable, ambiguity may lead individuals to adopt an "all-or-nothing" strategy: either being fully honest to avoid any negative consequence or fully dishonest as uncertainty about detection applies equally to all untrue statements. In this vein, it is plausible to assume that if an individual decides to cheat under ambiguity they will do so more likely to the maximum extent possible. Whether the distribution under ambiguity consists of more honest than dishonest behavior is also not entirely clear. Literature has shown that ambiguity may intensify individual risk preferences [Ghosh and Ray, 1997] and as most individuals are assumed to be risk-averse, this could result in a higher proportion of honest behavior. Conversely, ambiguity may also enable greater self-justification for dishonest behavior [e.g., Pittarello et al., 2015].

In summary, individual dishonest behavior is likely not only affected by the nature of the verification entity itself but also by whether machines operate the detection process under transparent or non-transparent rules. As there are convincing arguments for both more and less dishonest behavior under each rule type, we refrain from a directional prediction in formulating our second hypothesis:

Hypothesis 2: Human dishonest behavior will differ when their statements' truthfulness is verified by machines under transparent or undisclosed rules.

2.3 Experiment

We conducted a one-shot, incentivized laboratory experiment in which participants entered a prize draw with a potential payoff of up to €90. The final payoff depended on each participant's decision and the outcomes of up to two lotteries. Only one winner was drawn per session, in line with a random incentive system – a well-established approach in experimental economics that has been shown to produce similar behavior as under deterministic payoff schemes [Charness et al., 2016; Camerer and Hogarth, 1999; Bolle, 1990; Tversky and Kahneman, 1981].

2.3.1 Experimental Design

The experiment comprised two main parts: the **Choice Part** and the **Verification Part**.

In the **Choice Part**, illustrated in Figure 2, subjects drew exactly one card randomly from an urn containing 100 cards numbered between 1 and 6. Subsequently, they confidentially reported their drawn number via a computer interface. Importantly, the reported number – in conjunction with

Verification Part results – would later determine the prize payoff for one randomly selected winner, calculated as the reported number multiplied by €15 (payoff range: €15 to €90). This setup created the opportunity for subjects to increase their potential payoff by overreporting the drawn number. After submitting their report, participants completed a series of questionnaires (see Section 2.3.3), before the prize winner was determined.

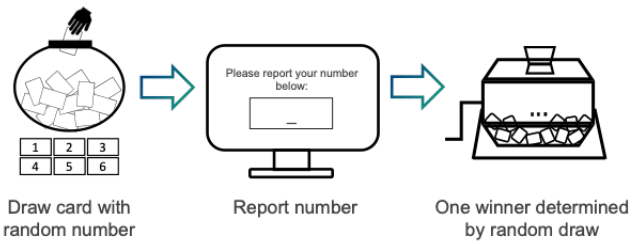


Figure 2: Overview of the Choice Part of the experiment (all participants)

In the **Verification Part**, the winner underwent a verification procedure comprising up to two lotteries:

- In *Lottery 1*, a number between 1 and 10 was randomly drawn. If this number was greater than the participant's reported number, no additional check occurred, and the full payoff (reported number × €15) was paid. If the number was less than or equal to the reported number, the participant's actual drawn card was checked.
- In the check, if the reported and actual numbers matched, the prize winner received the full payoff.
- If they mismatched, *Lottery 2* was triggered: An urn containing numbers from 1 up to the reported number was used to randomly draw one number. If this drawn number was less than or equal to the participant's actual number, the price winner still received the full payoff. Otherwise, the payoff was reduced to the actual number multiplied by €7.50 (payoff range: €7.50 to €37.50).

Thus, the verification procedure incorporated two central design features. First, the probability of a card check increased with the magnitude of the reported number – similar to materiality thresholds in accounting, where more conspicuous reports are subject to greater scrutiny. Second, the probability of punishment, conditional on being checked, increased with the discrepancy between the reported and actual number. This mechanism allowed subjects to potentially receive the full payoff despite overreporting, thereby mimicking discretionary tolerance in real-world verifications, where minor deviations may be overlooked while larger discrepancies are more likely to result in sanctions.

The structure of the Verification Part is illustrated in Figure 3.

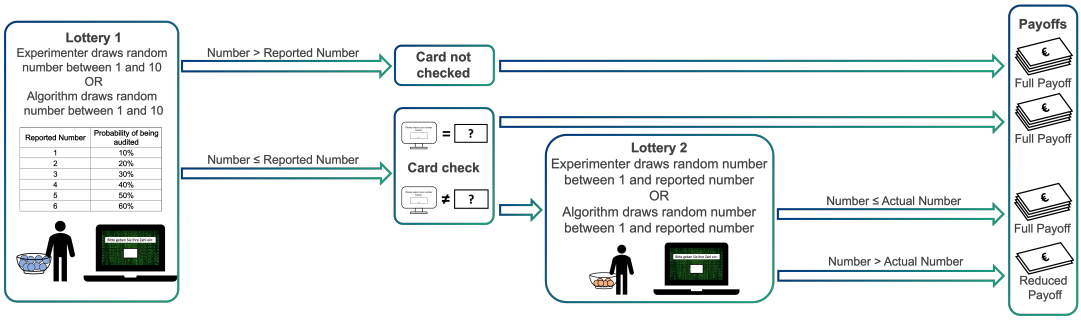


Figure 3: Overview of the Verification Part of the experiment (only prize draw winner)

Theoretically, for a risk-neutral decision-maker, the payoff-maximizing strategy is to always report a 6, regardless of the actual number drawn. The design of Lottery 2 ensured that cheating of equal magnitude carried identical punishment probabilities – for instance, overstating a 3 as a 5 was punished with the same likelihood as overstating a 2 as a 4. Meanwhile, the likelihood of punishment increased with the extent of the misreport: for example, if two subjects both reported a 6, the subject who actually drew a 2 faced a higher probability of being penalized than one who drew a 5. The formal derivation of this utility function is provided in Appendix A.1.

2.3.2 Treatment Conditions

We implemented four experimental treatments: **Human (H), Machine (M), Human Black Box (HB), and Machine Black Box (MB)**. In the Human treatment, the verification process was conducted by a human agent (i.e., the experimenter), whereas in the Machine treatment, it was executed by a computerized, rule-based algorithm. To emphasize the verification entity's role, both were visually represented using pictographs in the instructions (see Figure 21 in Appendix A.5). Procedures in the black box versions (HB and MB) mirrored their respective non-black box treatments (H and M), except that the verification rules were not disclosed to participants.

In the Human treatments, the lotteries of the verification part were physically implemented using numbered balls drawn from urns. In the Machine treatments, the process was simulated by a computer algorithm, with visual feedback (e.g., animations; see Figures 22 to 24 in Appendix A.5) provided to convey the impression of data processing. Critically, the underlying verification rules and their associated probabilities were held constant across all treatments; only the entity who conducted the process (human vs. machine) varied.

In the black box conditions, the exact same procedures were applied (verification rules and probabilities remained identical). However, subjects were only informed that a human or machine would decide whether a card check would occur and, in the case of a mismatch, whether the payoff would be reduced. To reflect this lack of procedural transparency, the verification steps were referred to as

"Decision 1" and "Decision 2" in the instructions.

In all treatments, while participants were informed the urn contained numbers 1 to 6, they were not told the actual distribution. The true composition of the urn was 95 cards displaying the number 2, while the numbers 1, 3, 4, 5, and 6 were each represented by a single card. This design ensured that most participants would draw a 2, allowing for individual-level analysis of dishonest behavior and increasing opportunities for overreporting. It would also largely prevent reduction of the sample size for the analysis due to subjects drawing a 6, which left them no opportunity to be dishonest. After each session, the remaining cards in the urn were counted to infer the actual distribution of numbers drawn. If all five non-2 cards remained, any report higher than 2 could be clearly identified as dishonest. If one or more of the five non-2 cards had been drawn, one observation with a report of a 6 would be randomly excluded from the dataset per card drawn, to obtain a conservative estimate of dishonest behavior.

This approach did not disadvantage any participant, as the distribution of cards was not disclosed in the instructions. The decision to equip the urn with a majority of cards numbered with a "2" instead of a "1" was made to avoid triggering "revenge cheating" (i.e., retaliation due to receiving the lowest possible draw) and to ensure participants faced a meaningful trade-off between honesty and financial gain. By drawing a "2" with the highest probability, truthful reporting would yield a €30 payoff for the prize winner, which is already substantial for an experiment participation of around 45-minutes, but could potentially be tripled through dishonest reporting.

2.3.3 Experimental Procedure

The experiment was conducted in December 2023 at the Business and Economic Research Laboratory (BaER-Lab, `www.baer-lab.org`) at Paderborn University and computerized using oTree [Chen et al., 2016]. Subjects were recruited via the online recruiting system ORSEE [Greiner, 2015] and were only allowed to participate in one session. In total, ten sessions were run (Human: 3, Machine: 3, Human Black Box: 2, Machine Black Box: 2). Each session lasted 30-45 minutes.

Participants were randomly assigned to individual computer workplaces in cubicles to ensure privacy and were instructed not to communicate during the session. After receiving written instructions (see Appendix A.2) and being given time to read them carefully, participants completed extensive comprehension checks to ensure a sufficient understanding of the experimental rules and payoff conditions. They could only proceed after answering all questions correctly. Consequently, subjects were, at least implicitly, aware of the opportunity to misreport before making any decisions in the experiment.

The Choice Part began once all subjects had successfully completed the comprehension checks. The experimenter moved from cubicle to cubicle, presenting an urn containing the number cards to

each subject. After the drawing process was completed, the experiment automatically advanced to the reporting screen, where subjects entered their reported number. To encourage thoughtful decision-making, participants were not subjected to any time limit.

After confirming their choice, subjects completed a series of questionnaires (see Appendix A.3). First, they were asked whether they generally preferred a human or a machine to perform the verification process. Second, subjects were asked which of the two entities they generally perceived as more error-prone and which as having greater discretion. Subsequently, subjects answered standardized questionnaires on affinity for technology interaction [Franke et al., 2018], attitudes toward ethical dilemmas [adapted from Blais and Weber, 2006], a pictorial measure of interpersonal closeness (adapted for inter-entity comparison) [Schubert and Otten, 2002, based on Aron et al. [1992]], the general risk preference measure by Dohmen et al. [2011], as well as demographic questions.

Once all questionnaires were completed, one prize winner was randomly selected using the cubicle numbers. Non-winning participants received a fixed payment of €7.50 in cash to compensate for their participation time[3] and were then dismissed.

The Verification Part was conducted privately with the winner to preserve anonymity and minimize social influence [Bolton et al., 2021][4]. The two lotteries were implemented based on the entity type of the respective treatment, following the procedure described in Section 2.3.1. The winner received their (full or reduced) payoff in cash, concluding the session.

2.4 Results

In total, one-hundred-seventy ($N = 170$) student subjects participated in the experiment. Of these, 48 were randomly assigned to the Human treatment (H), 41 to the Machine treatment (M), 43 to the Human Black Box treatment (HB), and 38 to the Machine Black Box treatment (MB) respectively. In the analysis, each subject constitutes one independent observation in the analysis. An overview of demographic characteristics is provided in Table 2. Participants were, on average, 22 years old, with ages ranging from 18 to 36. Women constituted 56% of the sample, and gender distribution did not differ significantly between treatments (Pearson $\chi^2(3) = 0.43, p = 0.935$). Multiple fields of study were represented, with Business Administration & Economics (56.5%) being the most common. The distribution of of fields of study did not differ significantly between treatments (Pearson $\chi^2(6) = 11.14, p = 0.084$)[5].

[3]This is three times the amount of the laboratory's usual show-up fee in experiments with individual performance-dependent incentives.

[4]While social image concerns toward the experimenter cannot be ruled out entirely, comparative statics ensure interpretability of treatment differences between groups.

[5]The complete dataset can be found in our OSF Repository.

Table 2: Demographic statistics

	H	M	HB	MB	Overall
Number of observations	48	41	43	38	170
Age					
Mean	21.8	21.8	21.9	22.5	22.0
Std. dev.	3.2	3.5	3.5	4.1	3.5
Gender (%)					
Female	54.2	58.5	58.1	52.6	55.9
Field of studies (%)					
Business Administration & Economics	56.3	68.3	58.1	42.1	56.5
Cultural Sciences	37.5	22.0	37.2	36.8	33.5
Natural Sciences	6.3	9.8	4.7	21.1	10.0

2.4.1 Dishonest Behavior

Similarly to Djawadi and Fahr [2015], our design enables a direct and relatively precise measurement of dishonest behavior – in contrast to prior experimental studies that infer dishonesty by comparing reported outcomes to theoretical distributions [see e.g., Abeler et al., 2014; Hao and Houser, 2008; Fischbacher and Föllmi-Heusi, 2013; Shalvi et al., 2011; Jacobsen and Piovesan, 2016] – by comparing the distribution of numbers drawn with the distribution of numbers reported. We use two dependent variables to measure cheating behavior: frequency and magnitude of overreporting, with the primary focus on the latter.

Figure 4 displays the frequency distributions of reported numbers by treatment. On average, subjects in the Human, Machine, and Human Black Box treatments reported numbers close to 3 (H: 3.06, M: 3.17, HB: 3.21), while subjects in the Machine Black Box treatment reported an average of 4.16. Reporting distributions differ significantly between groups (Pearson $\chi^2(15) = 33.07, p = 0.005$).

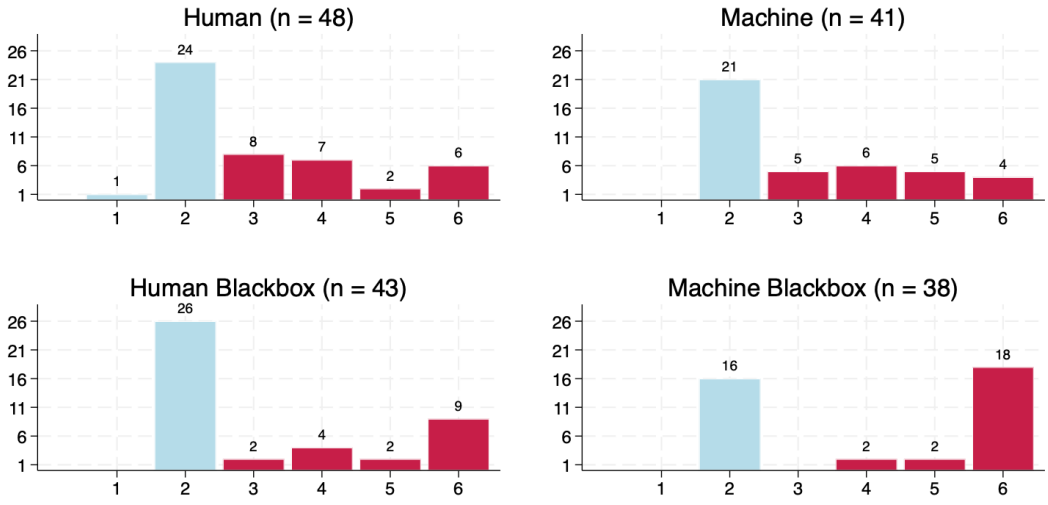


Figure 4: Frequency Distributions of Reported Numbers, by Treatment

In all treatments except the Human condition, no other numbers than 2 were drawn. In the Human treatment, the number 1 was drawn and accurately reported. Therefore, no exclusions of observations from the reported distributions were necessary, and any reported number above 2 can be interpreted directly as cheating.

In the non-black box groups, nearly half of the participants overreported: 23 out of 48 (47.9%) in the Human treatment and 20 out of 41 (48.7%) in the Machine treatment reported a higher number than they actually drew. Overreporting was less prevalent in the Human Black Box group (17 out of 43, or 39.5%), while the highest rate occurred in the Machine Black Box group (22 out of 38, or 57.9%). However, these differences in reporting rates are not statistically significant (Pearson $\chi^2(3) = 2.73, p = 0.435$).

Table 3: Summary statistics of cheating behavior by treatment

	H	**M**	**HB**	**MB**	**Overall**
Type of behavior (%)					
Honest	52.1	51.2	60.5	42.1	51.7
Partial cheating	35.4	39.0	18.6	10.5	26.5
Full cheating	12.5	9.7	20.9	47.4	21.8
Magnitude of cheating					
Mean	2.26	2.40	3.06	3.73	2.85
Median	2	2	4	4	3
Std. dev.	1.21	1.10	1.14	0.63	1.19

Note: Summary statistics of behavior type (relative frequencies) and cheating magnitude (among cheaters; absolute magnitude) by treatment. Instances of dishonest reporting: H: n = 23; M: n = 20; HM: n = 17; MB: n = 22.

Following conventions in related studies, we classify participants who overreported to the maximum extent possible (i.e., reporting a "6") as full cheaters, and those who overreported by a smaller margin as partial cheaters. Overall, the distribution of honest participants, partial cheaters, and full cheaters (see Table 3) differs significantly between treatments (Pearson $\chi^2(6) = 25.93, p < 0.0001$). In the non-black box groups, partial cheaters outnumber full cheaters. In the Human Black Box group, the proportions are roughly equal. In contrast, the Machine Black Box condition shows a substantially larger share of full cheaters, with partial cheaters being nearly absent. Notably, over half of participants were honest in the Human, Machine, and Human Black Box groups respectively, while the number of subjects who overreported to the maximum extent in the Machine Black Box group was higher than the number of honest subjects.

Regarding the magnitude of cheating, among cheaters, the average overreporting exceeded two numbers in all conditions, but was markedly higher in the black box groups. Consistently, the median magnitude of cheating was 2 in the non-black box treatments and 4 in the black box treatments. A Kruskal–Wallis equality-of-populations rank test with ties reveals a statistically significant difference

in cheating magnitude across groups (Pearson $\chi^2(15) = 21.64, p = 0.0001$). The Machine Black Box group not only shows the highest average cheating magnitude but also the lowest standard deviation, indicating more consistent and extreme overreporting, reflecting the group with the highest proportion of full liars.

Comparing magnitudes of cheating under transparent verification rules, we find no significant differences between the Human and Machine entity treatments (Mann–Whitney U-test: $|z| = 0.48, p = 0.6357$), as the average magnitude of cheating is only marginally higher in the Machine treatment than in the Human treatment. Under undisclosed rules, however, we observe a notable difference in cheating magnitude, as average overreporting is 0.7 higher in the Machine Black Box group than in the Human Black Box group – a difference that is statistically significant (Mann-Whitney U-test: $|z| = 2.09, p = 0.0442$). We therefore find partial support for **Hypothesis 1**, as the average magnitude of cheating differs by verification entity, but only under undisclosed verification rules.

Focusing on the machine groups under transparent and undisclosed verification rules, we observe a substantial increase in the average extent of overreporting – by approximately 1.3 – with the introduction of ambiguity about verification rules in the Machine Black Box group compared to the Machine group. The difference is highly statistically significant (Mann-Whitney U-test: $|z| = 4.03, p < 0.0001$). Therefore, we find support for **Hypothesis 2**: average magnitude of cheating toward a machine as verification entity differs between transparent and undisclosed processing rules, as ambiguity appears to lead to a higher magnitude of cheating. For comparison, overreporting toward a human as verification entity significantly increased by, on average, 0.8 from the Human to the Human Black Box (Mann-Whitney U-test: $|z| = 2.02, p = 0.0469$) [6].

To compare effect sizes, we calculate Cohen's d with bootstrapped standard errors (see Figure 33 in Appendix A.4). The entity effect is negligible in size under transparent verification rules ($d = -0.12$), while increasing to $d = -0.75$ under ambiguous rules, which can be classified as medium to large based on conventional benchmarks [Cohen, 1988]. Analogously, the effect of ambiguity in machine verification can be considered (very) large ($d = -1.50$).

2.4.2 Control Variables

The analysis of our questionnaire data provides strong support for the assumption that participants perceive humans as both more error-prone and more discretionary in their decision-making, as illustrated

[6] We conducted hypothesis testing based on the sub-sample of individuals who engaged in dishonest behavior, i.e., overreported their drawn number, as we argue that including honest reports would dilute the true extent of damage caused by cheating. Naturally the average magnitude of overreporting declines when these are incorporated (H: 1.1; M: 1.2; HB: 1.2; MB: 2.2). Nevertheless, key statistical results would remain robust: under undisclosed verification rules, the entity effect remains statistically significant (Mann–Whitney U-test: $|z| = 2.19, p = 0.0296$), as does the effect of ambiguity with a machine verifying the reports (Mann–Whitney U-test: $|z| = 2.28, p = 0.0214$), while still no significant difference is observed between entities under transparent rules (Mann–Whitney U-test: $|z| = 0.25, p = 0.8076$).

in Figures 5 and 6, Binomial tests for both variables yield results significantly different from 0.5 – which would indicate indifference – across all four treatment groups ($p < 0.0000$). Moreover, response distributions do not differ significantly between groups (error-proneness: Pearson $\chi^2(3) = 1.50, p = 0.681$; discretion: Pearson $\chi^2(3) = 1.26, p = 0.739$).

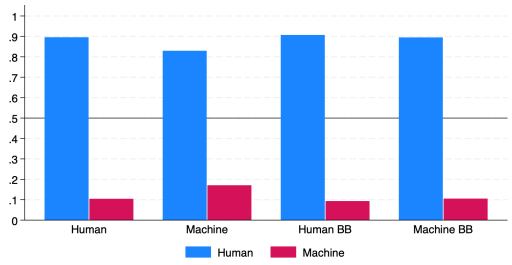


Figure 5: Perceived Error-proneness, by Treatment

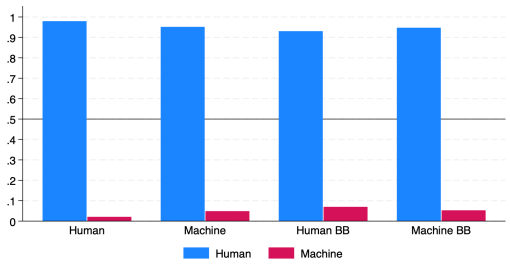


Figure 6: Perceived Discretion, by Treatment

Findings are less conclusive regarding participants' preferred entity for verifying the reports (see Figure 7). In both human treatment groups, participants tended to prefer a human as verification entity, whereas in the machine treatments, preferences leaned toward a machine as verification entity. However, in none of the groups did the distribution of preferences differ significantly from an even 50/50 split (see Table 32 in Appendix A.4 for Binomial test results by group). The apparent tendency to prefer the respective verification entity encountered during the experiment may reflect a default option effect [Johnson and Goldstein, 2003], as preferences were elicited post-experiment.

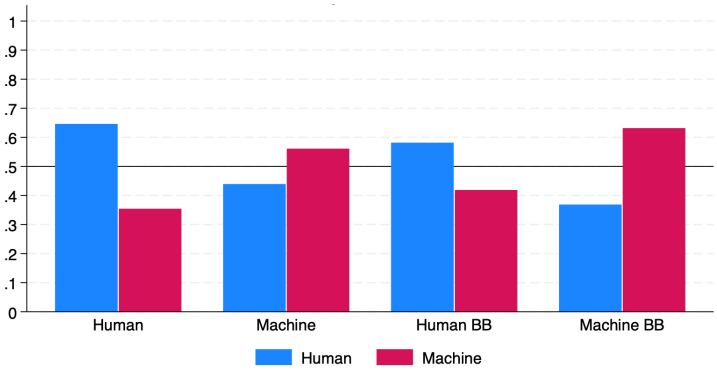


Figure 7: Stated Preference for Verification Entity, by Treatment

Furthermore, standardized questionnaire controls indicate that self-reported affinity for technology interaction, sensitivity to ethical dilemmas, perceived closeness to the verification entity, and stated risk preferences did not differ substantially across experimental groups as shown in Table 4[7].

[7]For pairwise treatment comparisons of cheating frequency and control variables see Table 30 in Appendix A.4

Table 4: Summary Statistics and Between-Group Comparison of Questionnaire Items

	H	**M**	**HB**	**MB**	**Total**	Kruskal-Wallis-H	
Number of observations	48	41	43	38	170	$\chi^2(3)$	p
Affinity to technology interaction	3.58	3.41	3.62	3.92	3.62	5.16	0.161
	(0.87)	(0.84)	(0.84)	(1.23)	(0.93)		
Ethical dilemma sensitivity	4.15	4.19	4.26	4.10	4.18	2.02	0.569
	(0.47)	(0.41)	(0.47)	(0.58)	(0.48)		
Interpersonal closeness	2.71	3.02	2.84	3.00	2.88	5.06	0.167
	(1.46)	(1.15)	(1.54)	(1.23)	(1.36)		
Risk preferences	5.77	6.22	5.77	6.03	5.94	1.37	0.713
	(2.15)	(2.24)	(2.16)	(2.11)	(2.15)		

Note: Summary statistics for affinity to technology interaction (6-point scale), sensitivity towards ethical dilemmas (5-point scale), perceived closeness towards the verification entity (7-point scale), and self-reported risk preferences (11-point scale). Standard deviations are reported in parenthesis. Kruskal-Wallis-H reports p-values for Kruskal-Wallis H-tests with ties between experimental groups.

Across all subjects, those who overreported and thus cheated reported a significantly higher willingness to take risks (Mann–Whitney U-test: $|z| = 2.62, p = 0.0085$). On average, cheaters indicated a general risk tendency of 6.4 (median: 7) on an 11-point scale, compared to 5.5 (median: 5.5) among honest participants. Also, the willingness to take risks was significantly positively correlated with the magnitude of cheating (Spearman's $\rho = 0.355, p = 0.0011$).

Also, gender differences were evident: women cheated significantly less frequently than men (Pearson $\chi^2(1) = 13.36, p < 0.0001$), with 35.8% of female and 64.0% of male participants overstating their drawn number. However, the magnitude of cheating did not differ significantly between genders (Mann–Whitney U-test: $|z| = 0.67, p = 0.5032$).

The other demographic and control variables did not differ significantly between honest and dishonest participants, nor were they significantly associated with the extent of cheating (see Table 31 in Appendix A.4).

2.4.3 Regression Analysis

In addition to our non-parametric analysis, we conduct multivariate regression analysis to gain a deeper understanding of the relationship between cheating behavior and its potential determinants. Based on the sub-sample of individuals who cheated ($n = 82$), we examined the factors influencing the extent to which participants overstated their drawn number. Specifically, we regressed the magnitude of cheating on the type of verification entity and the ambiguity level of verification rules, along with demographic, control, and entity-perception variables. Table 5 presents the results of the multivariate OLS regression, comparing multiple model specifications.

The baseline model (Column 1) includes only treatment indicators as independent variables, while subsequent models add demographic variables (Column 2), control variables (Column 3), and dummy variables indicating matches between the assigned verification entity and participants' stated entity preferences, perceptions of error-proneness, and perceived discretion (Column 4) respectively. All available variables are included in the full model (Column 5). For the sake of completeness, Tables 34 and 36 in Appendix A.4 present a linear probability model and marginal effects from a logistic regression estimating the independent variables' influence on the likelihood of cheating across the full sample. Both used the same model specifications as those employed in the regression for cheating magnitude. These robustness checks yield results consistent with our non-parametric analysis, with gender and general risk preferences emerging as the only statistically significant and substantively meaningful predictors of likelihood to cheat. For instance, being female is associated with a 30.7 percentage-point lower probability of overreporting.

Table 5: OLS Regression for Magnitude of Cheating

	Dependent variable: Magnitude of cheating				
	(1)	(2)	(3)	(4)	(5)
Intercept	2.261***	1.480*	-1.069	2.218***	-0.835
	(0.253)	(0.675)	(1.068)	(0.616)	(1.564)
Treatment					
Machine	0.139	0.067	2.500**	0.245	1.823
	(0.353)	(0.371)	(0.795)	(0.615)	(0.924)
Human Black Box	0.798*	0.625	0.478	0.816*	0.486
	(0.375)	(0.364)	(0.364)	(0.364)	(0.352)
Machine Black Box	1.466***	1.640***	3.957***	1.600**	3.520***
	(0.287)	(0.252)	(0.814)	(0.580)	(0.944)
Age		0.054			0.041
		(0.030)			(0.032)
Female		-0.391			-0.357
		(0.218)			(0.221)
Field of Study					
Cultural & social studies		−0.530*			-0.325
		(0.251)			(0.244)
Natural science		−0.939**			-0.369
		(0.303)			(0.372)
Risk			0.138*		0.149*
			(0.066)		(0.056)
Ethical sensitivity			0.151		0.159
			(0.227)		(0.229)
Closeness			0.080		0.086
			(0.064)		(0.071)
Verification by machine # ATI					
0			0.466*		0.183
			(0.192)		(0.208)
1			-0.225		−0.334*
			(0.132)		(0.147)
Verification by preferred entity				-0.242	-0.051
				(0.227)	(0.219)
Verification by more error-prone entity				0.812*	0.806**
				(0.308)	(0.295)
Verification by higher discretion entity				-0.517	-0.624
				(0.535)	(0.560)
F-test	13.31***	14.59***	9.64***	10.87***	10.53***
R^2	0.2600	0.3712	0.4342	0.3454	0.5527
Adj. R^2	0.2615	0.3117	0.3722	0.2931	0.4510
N	82	82	82	82	82

Note: Coefficients estimated using robust standard errors, standard errors in parentheses; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Model specifications: (1) treatment variables only, (2) including demographics, (3) including control variables, (4) including entity perceptions, (5) full model.

Among the model specifications, the full model (Column 5) yields the highest coefficient of determination ($R^2 = 0.5527$), which is substantially high for studies based on observational data on human behavior. Accordingly, the model explains a considerable share of the variation in cheating magnitude. The adjusted R^2 is about 10 percentage points lower, reflecting the inclusion of numerous explanatory variables. Consequently, our interpretation of results focuses primarily on this specification.

Consistent with the non-parametric findings, the Machine Black Box treatment stands out: its coefficient is substantially larger – indicating that overreports are, on average, 3.5 units higher – and significantly different from that of the Human group, which serves as the reference category in the regression. In contrast, the coefficients for the Machine and Human Black Box treatments are smaller in magnitude and not significantly different from the Human group. This pattern suggests that it is specifically the combination of audit ambiguity and a machine auditor that drives the increase in dishonest reporting.

While being male was a major predictor of the likelihood to cheat, gender does not significantly affect the magnitude of cheating. In contrast, individuals' risk preferences are significantly related to both the decision to cheat and extent of cheating. Specifically, a one-point increase in self-reported risk willingness to take risk is associated with an average increase of 0.15 in the magnitude of overreporting. Though modest in size, this effect accumulates across the 11-point scale. As anticipated from the non-parametric analysis, regression coefficients for other demographic and control variables – ethical sensitivity, perceived closeness to the verification entity, age, and field of study – are neither statistically significant nor meaningful in size.

Notably, an individual's affinity for technology interaction (ATI) appears to be associated with reduced cheating magnitude, but only when the verification is conducted by an algorithm. In these cases, each one-point increase in ATI (on a 6-point scale) corresponds to an average decrease of 0.33 in the magnitude of cheating. This suggests that individuals who feel more comfortable with technology tend to cheat less under machine verification, potentially due to better understanding an algorithm's capabilities, even though they do not report completely honestly. No comparable effect is observed under human verification, which appears intuitive as there is no connection between reporting and technology for them.

Regarding the discussed psychological drivers of cheating, only the perception of the verification entity as more error-prone appears to be consequential. When the assigned auditor matches the participant's perception of being the more error-prone entity, while having been irrelevant for the likelihood to cheat, the magnitude of cheating increases by approximately 0.8. By contrast, whether the verification entity is perceived as having greater discretion does not have a significant impact on cheating magnitude.

2.5 Discussion and Conclusion

Human-machine interactions become increasingly pervasive in daily life and professional contexts, motivating research to examine how human behavior changes when individuals interact with machines rather than other humans. While most of the existing literature focuses on human perceptions and actions toward algorithmic systems in advisory roles, our study examines a different yet equally important human-machine setting in which machines can detect untrue statements of humans and penalize their fraudulent reporting. We incorporate elements of risk and uncertainty into the die-roll paradigm by Fischbacher and Föllmi-Heusi [2013] and design four experimental conditions varying the verification entity (human versus machines) and the transparency of processing rules (transparent versus ambiguous) to detect and sanction dishonest behavior. The experimental design involved a clearly quantifiable reporting task in which participants could increase their earnings by overreporting the actual outcome of the die-roll, while facing either specified or unknown risks of detection and punishment. Unlike many earlier studies where deception carried no consequences for the individual, our design reflects realistic decision environments where risk preferences matter, payoff incentives are substantial, and higher reported values face greater scrutiny.

Cheating was observed – at relatively high rates between roughly 40% and 60% – across all four experimental conditions. In each treatment, we observed the full spectrum of behavior: complete honesty, partial cheating, and full cheating. Under transparent processing rules, we do not find a behavioral difference in cheating magnitudes between humans and machines as verification entities. This finding is consistent with literature which argues that behavioral differences may arise if functionally equivalent actions performed by humans and machines are perceived differently [e.g., Bigman and Gray, 2018; Bogert et al., 2021]. Under transparent rules, such perceptual differences appear to be largely neutralized. When individuals know exactly the verification procedure and understand that the verification entity is bound to that procedure, potential differences in social image concerns or moral considerations that might otherwise differentiate human-machine interactions are minimized. Consequently, participants' behavior converges toward a rational response to the underlying risk-reward structure, regardless of whether statement verification is conducted by human or algorithmic agents. When detection rules are not known to individuals and are thus ambiguous, significant behavioral differences in cheating magnitude emerge. Most notably, human dishonesty differs between the "Machine" and the "Machine Black Box" conditions, highlighting the pivotal role of algorithmic opacity or the black box nature of algorithms and AI systems. We hereby find strong evidence of higher average cheating magnitudes when machines verify under ambiguous rather than transparent rules. Specifically, the behavioral pattern under machine ambiguity exhibits increased polarization, with participants more likely

to engage in either complete honesty or maximal dishonesty, rather than partial cheating. The fact that in aggregation these average cheating magnitudes are significantly higher than in the transparent condition indicates that ambiguity facilitates greater justification for dishonest behavior. We observe a similar trend of behavioral differences in conditions where a human serves as the verification entity but not to the same extent as with machines. Specifically, we find that average cheating magnitude in the "Machine Black Box" treatment is significantly higher than in the "Human Black Box" treatment. In line with prior work by Cohn et al. [2022] and Biener and Waeber [2024] where their experimental designs come nearest to our black box conditions, differing social image concerns toward humans and machines as verification entities could explain the observed treatment differences. This suggests that overreporting to a human is more readily perceived as morally questionable, whereas overreporting to a machine may be more likely construed as engaging in morally neutral gambling behavior. This reasoning also helps explain why instances of cheating decrease in the "Human Black Box" treatment compared to the "Human" treatment, while increasing in the "Machine Black Box" treatment relative to its transparent counterpart. Under ambiguous conditions, individuals appear to suspend or attenuate internalized norms of honesty when interacting with machines. This behavior could be further interpreted through the lens of self-serving belief distortion [Bicchieri et al., 2023], where individuals strategically reinterpret the ethical dimensions of their actions when circumstances permit moral flexibility. The combination of machine verification and algorithmic ambiguity may create exactly that condition which facilitates such ethical re-framing, enabling individuals to justify dishonest behavior that they might otherwise consider morally problematic. In summary, these findings support our entity type hypothesis partially: behavioral differences in dishonesty between humans and machines as verification entities do not emerge in general, but specifically under conditions of ambiguous detection rules.

Overall, cheating rates in our experiment appear relatively high compared to related studies, with no clear evidence of a general "preference for truth-telling" [Abeler et al., 2019]. This may be attributed to the explicit risk component in our design. Unlike other studies where cheating involves implicitly violating the rules of the game and the social norm of honesty, our task explicitly included the possibility of sanctions, thereby making participants consciously aware of both the opportunity to cheat and its potential consequences. We do not view this as problematic in terms of potential experimenter demand effects, as the research objective focused on comparative rather than absolute levels of dishonesty. Any upward bias in overall cheating due to heightened salience of sanctioning dishonest behavior would not systematically affect between-group comparisons. Furthermore, we carefully designed the instructions to be neutral and avoided language with ethical connotations such as "lying", "cheating", or "punishment" (see Appendix A.2).

However, the results and implications of our study should be interpreted with caution, given its methodological and contextual limitations. First, the number of participants per treatment group is relatively modest. This means that sub-samples of cheaters are even smaller, which may limit the statistical power of our analysis (see Figure 33 in Appendix A.4). Consequently, findings based on medium effect sizes and p-values near the 0.05 threshold should be interpreted cautiously. Nonetheless, effects related to ambiguity ($d > 1$) and the apparent absence of entity effects under transparent detection rules are sufficiently distinct to support clearer conclusions.

Second, despite the machine verification procedure being framed as algorithmic, the experimenter remained involved in its administration. In particular, the drawn number was still checked by a human. While this setup does not entirely eliminate potential social image concerns toward the experimenter, the comparative statics should preserve the interpretability of between-group differences. Meanwhile, perceptions of anthropomorphism toward the algorithm should be negligible, as subjects visibly interacted with a computer interface with no human-like features (see Appendix A.5). Furthermore, our study explicitly referred to the machine verification entity as an "algorithm". Therefore, extrapolation of our results to contexts involving broader concepts like "artificial intelligence" should be done with care. AI systems may be perceived as more autonomous or human-like than basic algorithms, potentially influencing behavior differently by invoking greater expectations of discretion or intentionality.

Third, the Verification Part of our experiment can be viewed as a compound lottery, a design feature that has been subject to discussion in elicitation literature [see e.g. Starmer and Sugden, 1991; Harrison et al., 2015]. However, our design requires subjects to make only a single consequential decision, aligning with how individuals are typically found to approach compound lotteries [Holt, 1986]. If the lottery design influences behavior at all, it is likely to do so by discouraging cheating due to incomplete understanding of the consequences – and could only do so in the transparent treatments, as the probabilistic structure of verification was undisclosed in the ambiguous conditions. Nonetheless, cheating rates in all four treatments can be considered medium to high compared to related studies. Furthermore, we preemptively addressed potential misunderstandings of the Verification Part by including a step-wise graphic illustration in the instructions, and requiring subjects to answer seven multiple-choice comprehension questions correctly before advancing to the Choice Part. Subjects were not informed which answers needed to be corrected if they erred, ensuring genuine understanding rather than trial-and-error guessing.

Finally, while internal validity appears relatively strong – a substantial part of regression model variation is explained by the covariates included, and high monetary incentives should largely neutralize outside preferences in line with induced value theory [Smith, 1976] – questions regarding external validity remain. Specifically, our design assumes an equidistant likelihood of punishment for equal

magnitudes of cheating, which may not reflect real-world audit procedures. However, we consider this a mere mathematical design feature of inferior relevance which was necessary to maintain comparability with other experimental cheating studies. Moreover, in our experiment, punishment was applied within a gain frame: even prize draw winners that were detected and punished exited the experiment with a positive net payoff. In real-world settings, penalties outweigh gains and result in actual losses – conditions that present significant methodological challenges for experimental replication. From a practical standpoint, while real-world verifications or audits typically do not operate under undisclosed or ambiguous rules due to legal constraints, perceived ambiguity often exists nonetheless, particularly among non-experts facing for example complex tax laws and legal regulations. Such perceived opacity may effectively replicate in practice the black box experience observed in our experimental conditions.

Future research could build on the aforementioned distinction of the terms "algorithm" and "AI" by directly examining interactions with AI-based systems, rather than simpler algorithmic tools. Beyond this, it would be valuable to replicate and extend our findings across more diverse participant cohorts. While we identify plausible relationships between gender and risk preferences with dishonest behavior, additional individual characteristics and underlying motives may serve as important determinants of dishonest behavior. For instance, previous studies have shown that older individuals and non-students generally exhibit lower levels of dishonest behavior compared to student samples [Djawadi and Fahr, 2015]. Similarly, domain experts tend to have different attitudes toward and behaviors in response to algorithmic decision-making than the general public [Jussupow et al., 2020]. Even though participants judged the human verification entity to exhibit more discretion, this perception appears to have played a secondary role in reporting decisions. In contrast, perceiving the verification entity as error-prone was found to increase the average magnitude of overreporting. However, this mainly applies to those conditions with a human auditor, as humans are nearly universally perceived as the more error-prone entity. Corroborating evidence from future research would be valuable in clarifying the roles of these factors as behavioral motivations in this particular human-machine interaction context. Similarly, given that participants' stated preferences indicated indifference about which entity should verify their reports, it would be interesting to examine whether this translates into actual behavior when partipants can select the verification entity under either transparent or ambiguous rules. In this regard, future research could test whether the findings by Cohn et al. [2022] can be replicated, namely that participants who intend to be dishonest select machine verification when processing rules are undisclosed. Moreover, future studies might consider adopting a double-blind payment procedure, such as that used by Fischbacher and Föllmi-Heusi [2013], to fully remove any residual human involvement in the machine verification process. Lastly, alternative incentive structures could be explored. For example, awarding smaller monetary prizes to multiple winners rather than a single large prize may

produce different motivations and cheating dynamics, offering further insight into the role of stakes and competition in dishonest behavior [Kajackaite and Gneezy, 2017; Martinelli et al., 2018; Rahwan et al., 2018].

Nevertheless our study carries important practical implications. When machines are planned as verification entities, we recommend that practitioners and policymakers prioritize addressing the black box problem by enhancing procedural transparency, i.e. "opening the black box" [Litterscheidt and Streich, 2020]. The combination of ambiguous rules and machine verification clearly drives up the magnitude of cheating and thus the related economic damage. While transparency alone may not eliminate dishonest behavior, a lack of transparency is likely to exacerbate it significantly. Given that our results suggest that the magnitude of cheating under ambiguity is lower when a human is involved, automating detection processes in such settings could unintentionally increase the impact of dishonest behavior. These findings therefore cast skepticism on the expectations of authorities, such as tax agencies, that automation may produce deterrence effects simply because machines can better identify suspicious patterns in tax reports. Rather, in contexts where rule interpretation is complex or ambiguous, it may be advisable to revert automated (verification or auditing) processes back to humans, provided that the cost of human employment is offset by the averted damage from dishonest behavior. Beyond the binary perspective of our experiment, hybrid solutions such as human-in-the-loop process designs, may offer valuable alternatives for ostensibly routine tasks that hold large damage potential in exceptional cases. For instance, AI can be used to improve efficiency in insurance claim processing and fraud detection by identifying inconsistencies or suspicious patterns in claim submissions, which are then forwarded for further human assessment and final decision-making [Komperla, 2023].

Conversely, when processing rules are transparent, algorithmic verifications may offer a viable and cost-efficient alternative without further sacrificing behavioral integrity. In such cases, the identity of the verification entity – human or machine – appears to have no meaningful effect on cheating behavior in terms of either frequency or magnitude. Natural areas of application include financial and tax audits, where algorithmic automation offers great potential for efficiency improvements [Bakumenko and Elragal, 2022; Li et al., 2025]. These systems are already used to determine audit targets, with researchers working to increase purposive selection and algorithmic fairness [Black et al., 2022]. For example, in some domains, such as tax administration, policy debates have emerged around requiring tax agencies to disclose their algorithmic procedures and inform taxpayers subjected to severe audits about the reasons for selection, thereby providing grounds for legal challenge [Faúndez-Ugalde et al., 2020].

However, our findings may be extended to all kinds of compliance, monitoring, and verification processes that hold potential for both automation and dishonest human behavior. For example, in set-

tings where electronic surveillance are installed to monitor human conduct, these systems are perceived more negatively than human surveillance systems [Schlund and Zitek, 2024]. While monitoring and surveillance are inherently unwelcome, ensuring that electronic surveillance systems are not perceived more negatively than human alternatives serves the interests of authorities and organizations. Empirical evidence suggests that electronic surveillance may trigger psychological reactance, a motivational state of resistance towards perceived restrictions on behavioral freedom, which frequently manifests in deviant behavior. For example, Yost et al. [2019] find that electronic surveillance in organizations elicits reactance that correlates with increased employee intentions to engage in counterproductive workplace behaviors. Based on our results, one approach to mitigate this perceptual gap may be enhancing transparency in monitoring rules and procedures so that individuals view the electronic system as substitute for, rather than intensification of, human surveillance. In this regard, automated solutions can be implemented such that the benefits of reduced human labor costs are not offset by increased costs arising from more dishonest or counterproductive workplace behavior.

Chapter 3:

"Does the involvement of domain experts in the AI training affect their AI perception and AI adherence? An experiment in the context of industrial AutoML applications"

Does the involvement of domain experts in the AI training affect their AI perception and AI adherence? An experiment in the context of industrial AutoML applications

Anastasia Lebedeva*, Marius Protte*,†, Dirk van Straaten*, René Fahr*

Abstract

AutoML is a promising field of Machine Learning (ML) that is supposed to bring the advantages of artificial intelligence to a wide range of organizations in plentiful domains, by automating the process of ML-model creation without requiring prior knowledge in data science or programming. However, AutoML often appears to users as a black-box model created in a black-box process, negatively impacting users' trust. Additionally, AutoML users are often experts in their respective domains (physicians, engineers, etc.), which are commonly observed to exhibit stronger algorithm aversion than lay people, i.e., having more difficulties trusting and relying on AI recommendations. User non-adherence to AutoML may have high-cost consequences, resulting in inefficient decisions and mitigating the overall progress in the AutoML field. Therefore, we investigate how domain experts' adherence to AutoML recommendations can be fostered. As involvement of users in product creation processes was shown to positively affect their attitudes towards the product in multiple contexts, we argue that involving domain experts in AutoML-model creation processes may increase their trust and adherence to AutoML. We conduct an experimental laboratory study, in which subjects act as expert engineers and need to foresee machine malfunctions, while being advised by an AutoML-model. We apply three treatments – zero, passive & active involvement – to investigate our hypothesis. We observe that higher involvement leads to a higher perceived influence on the AutoML model and a higher perceived understanding of its functionality. However, these perceptions are not reflected in the actual behavior – subjects across all groups demonstrate similar AI adherence. We suspect that subjects' perceived expertise, which is equally strong across all groups, may overrule any effect of involvement.

JEL Classification: C91, D24, L23, O14, O33

Keywords: algorithm aversion; user involvement; AutoML; laboratory experiment

*Paderborn University, Heinz-Nixdorf-Institute, Fürstenallee 11, 33102 Paderborn

†Corresponding author, marius.protte@upb.de

This article was published in *Advances in Information and Communication: Proceedings of the Future of Information and Communication Conference 2024* under the title *"Involvement of domain experts in the AI training does not affect adherence – An AutoML study"*, DOI: https://doi.org/10.1007/978-3-031-53960-2_13. The present version deviates from the published version in minor textual and organizational changes.

This research was funded by the Deutsche Forschungsgemeinschaft within the "SFB 901: On-The-Fly (OTF) Computing – Individualised IT-Services in Dynamic Markets" program (160364472).

3.1 Introduction

A growing number of organizations are willing to apply Machine Learning (ML) to generate value out of their data. However, building and deploying an ML-Model is generally associated with high investments and requires involvement of data scientists and other ML-professionals, who are rare on the labor market. These challenges prevent medium and small organizations from adopting ML and mitigate the overall spread of this technology. One approach to tackle this problem is known as Automated Machine Learning (AutoML) [Singh and Joshi, 2022]. AutoML is a paradigm which seeks to automate the complex development process behind Machine Learning and to make it more accessible. The ultimate vision behind this technology is to enable domain experts – users without data-science skills, but with expertise in the application domain – to build and deploy ML-models for their specific real-world problems [Karmaker et al., 2020].

Though in the recent years the AutoML technology has demonstrated remarkable progress, it is still a long way till domain experts can autonomously build, apply, and interpret AutoML-models. Currently, from the user perspective, AutoML is a black-box process which leads to a black-box solution, both difficult or even impossible to interpret without data science skills. Such lack of transparency has multiply negative implications, one of them being strong mitigation of users' trust [Zöller et al., 2022]. In general, a phenomenon of discounting an algorithm judgment in favor of the own or other humans' judgment, irrespective of algorithm performance, is known as algorithm aversion [Mahmud et al., 2022]. Research has demonstrated that people are averse towards algorithms, even when they know about algorithm superior performance [Dietvorst et al., 2015]. The trust for AutoML can be additionally negatively impacted by the fact that most AutoML-users are experts in their respective domains – engineers, physicians etc. According to Logg et al. [2019], experts rely less on algorithms in their decision-making compared to lay users, even at the cost of decision accuracy. Filiz et al. [2021] show that, paradoxically, people demonstrate even stronger algorithm aversion when decision stakes are higher – they strongly prefer a less accurate human advice over a more accurate algorithm advice, if a situation has critical impact on others. AutoML as a technology will eventually bring Machine Learning into a broad range of domains, high-risk domains like healthcare or industrial production among them. Consequently, algorithm aversion in AutoML-context may not only reduce users' own utility from AutoML but lead to high-stake failures. Therefore, it is a crucial question, how trust in AutoML can be fostered.

A comprehensive literature review by Jussupow et al. [2020] summarizes that human involvement at different stages of algorithm life cycle mitigates algorithm aversion. In a broader scope of information systems research, a meta-analysis of 82 empirical studies by He and King [2008] reveals that user

participation has a significant positive effect on attitudinal and behavioral outcomes (user satisfaction, use intention and system use). Behavioral sciences also offer support for the notion that co-creation positively influences perception of results. For example, the widely known IKEA-effect, introduced by Norton et al. [2012], describes a phenomenon when co-creation leads to higher valuation of products by consumers. Psychological ownership for the result of co-creation is shown to be one of major mechanisms underlying this behavior [Sarstedt et al., 2016]. In the AutoML-context, domain experts can act not only as users, but also as model contributors or even creators. Through their contribution, domain experts gain (to some extent) control over the resulting model. Dietvorst et al. [2018] show that an opportunity to exhibit control over an algorithm (even slightly) reduces algorithm aversion. Besides, involvement during model training allows domain experts to insert their knowledge into the AutoML-model. Kawaguchi [2021] demonstrate that sales workers become more willing to follow algorithmic advice once their expertise is integrated into the algorithm's forecasting. Additionally, by participating in the model creation process, domain experts gain better understanding about functionality and inputs of an AutoML-model. According to Yeomans et al. [2019], information about algorithm functionality improves users' understanding of the algorithm and reduces algorithm aversion. Finally, Jago [2023] demonstrates that highlighting human involvement in model training can have a positive effect on user perception.

Provided theoretical insights from different scientific fields demonstrate that involvement of domain experts during model creation may have a value of its own: Enable domain experts to influence the model, better understand it, develop a feeling of psychological ownership towards the model and eventually foster trust in AutoML and enhance algorithm adherence. To the best of our knowledge, current AutoML-research treats involvement of domain experts as a technical measure to insert domain-specific information into the model: The value of such involvement is determined by the value of inserted information, its carrier being simply a source. Involving domain experts has a cost of time and effort, so a rational decision-maker would compare the cost with the anticipated benefit (e.g., higher accuracy of the model) to decide whether a model-training should happen with or without expert involvement. In general, it is a decision which a domain expert herself might be confronted with – should she invest time, sharing her knowledge or is it sufficient to train the model based only on data? We argue that involvement of domain experts in model creation has a value beyond the technical improvement of the model and that this value must be scientifically determined in order to be considered by practitioners and decision-makers. We make a first step in this direction by empirically investigating the effect of user involvement in AutoML-context. Explicitly, we pose the following research question:

How does involvement of domain experts during model training affect their adherence to model advice during its deployment, given that the model accuracy does not change through involvement?

To answer our research question, we state following hypothesis:

Hypothesis 1a: Expert individuals who were involved in model training are more likely to adhere to the model advice compared to individuals who were not involved.

Hypothesis 1b: Higher grade of involvement in model training is associated with higher adherence to model advice.

Our experimental study is designed as a simplified real-world situation, modelling a predictive maintenance problem in industrial context. Within this scenario, subjects play a role of engineers, responsible for the maintenance of a manufacturing machine. They need to assess the probability of a machine malfunction and take a maintenance decision, while being advised by an AI. During the experiment, subjects gradually acquire scenario-relevant knowledge and become domain experts within the scenario. For the purpose of our inquiry, we vary the grade of subjects' involvement during model training and implement three treatment groups: zero, passive and active involvement. Following our hypothesis, we measure, how often subjects change their decision in favor of the AI advice and compare the advice adherence between the treatment groups. We implement monetary incentives for maintenance performance in order to enhance the internal validity of our study and to elicit subjects' true preferences [Falk and Heckman, 2022].

With our paper we seek to make several contributions, valuable both for researchers and practitioners. Firstly, we empirically examine the impact of user involvement during model training on algorithm adherence. Secondly, we add to the literature on algorithm adherence among expert users. Thirdly, we add to the literature on AutoML by empirically studying domain experts as contributors and as users of an AutoML-model, holistically addressing behavioral aspects of their role. Practitioners can benefit from our results by reassessing the role of domain experts and developing measures to better incorporate them into the AutoML-process. Additionally, they can directly influence the reliance on algorithm recommendations by varying the grade of domain expert involvement.

3.2 Method

We conduct an incentivized laboratory experiment, modelling a real-world situation on a production plant, in which a decision-maker must weigh a cost of a maintenance break with a cost of a potential machine malfunction. Throughout the experiment, subjects are asked to take a role of an engineer –

a domain expert for a manufacturing machine. In their decisions, subjects are confronted with an AI that predicts the probability of machine malfunctions and provides recommendations on whether to conduct a maintenance or not.

3.2.1 Experimental Design

In the beginning of the experiment, subjects are given all necessary knowledge and skill to play their role as a domain expert within the experimental scenario. Further, subjects are being passively, actively, or not involved in the AI training based on the applied treatment. In the main part of the experiment, subjects are asked to assess the likelihood of a fictitious machine breaking down and to decide, whether to maintain the machine or not in multiple rounds. Conducting maintenance eliminates the possibility of a malfunction and costs a half of the payoff. If no maintenance is conducted, subjects either receive a full payoff if no malfunction occurs and a zero payoff, if malfunction takes place. The actual probability of a malfunction depends on parameters, which are not known to subjects but can be assessed based on available data. The design follows the judge-advisor principle – in every round subjects first make their own assessment and maintenance decision, then learn about the AI-assessment and decision recommendation and take a final decision, to maintain a machine or not. To prevent learning effects, subjects receive no feedback in the end of a round and learn their total payoff only after they have completed all experimental rounds. In the end of the experiment, subjects are asked to fill in a questionnaire.

As mentioned above, in advance to the main part of the experiment, subjects learn to be domain experts within the experimental scenario. For this purpose, they are provided with information on how to assess the probability of a machine malfunction. For the sake of simplicity and comprehensibility, this information is reduced to the minimum required by the scenario. In our scenario, the actual probability of a malfunction is unknown to subjects, but can be estimated using three indicators: Temperature (of the motor), Speed (of the conveyor belt) and Workload (of the production line). The indicators can take unit-free values between 0 and 100. Each of the indicators has its unique optimal range, with a malfunction being less likely when an indicator lies within the optimal range. The more of the indicators lie inside their respective optimal ranges, the less the overall probability of a malfunction. For the sake of simplicity, the likelihood of malfunctions follows a four-step ordinal scale: "very unlikely", "unlikely", "likely" and "very likely" [8]. Subjects are provided with a heuristic rule: If all three indicators are located within their optimal ranges, a malfunction is considered "very unlikely". For each indicator located outside its optimal range, a malfunction becomes one "step" more likely – i.e., a malfunction is "very likely", when all indicators are outside their optimal ranges

[8] These correspond to underlying probabilities of 5%, 35%, 65% and 95% for a malfunction occurring this round (unknown to subjects).

(for an example, see Figure 8).

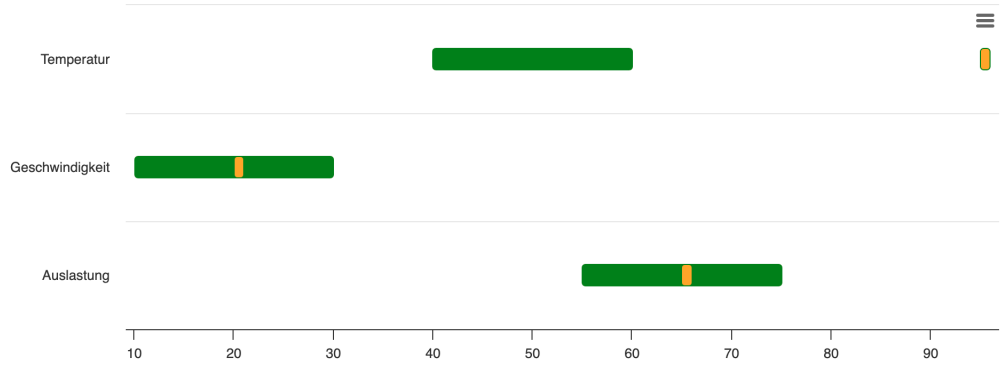


Figure 8: Optimal Ranges (green bars) and Values (orange dots) of the Three Indicators (notation in German). In this example, Temperature is located outside its optimal range and Speed and Workload are located inside their respective optimal ranges. Accordingly, a malfunction would be considered "unlikely".

To model the ambiguity of a real-world situation, the exact optimal ranges remain unknown to subjects. Instead, they need to estimate them based on historical data. subjects' individual assessments of optimal ranges are called "acceptable ranges". Acceptable ranges serve subjects as an approximation of the (unknown) optimal ranges in the heuristic rule described above. To set their acceptable ranges, subjects receive – for each indicator separately – a graphic distribution of past machine states in correspondence with indicator values. subjects are asked to choose a range of indicator values, which they consider acceptable based on the number of past malfunctions within this range. The distribution for each indicator is unique and contains sixty data points, thirty with a malfunction and thirty without (see Figure 9)[9].

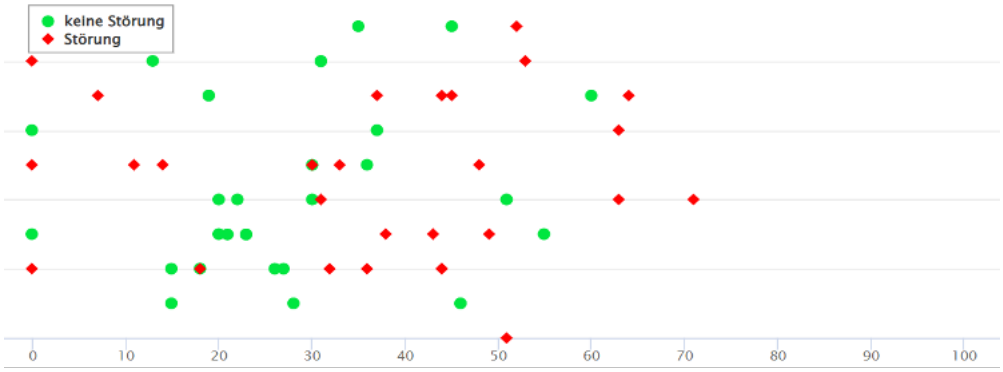


Figure 9: Example Distribution for the Indicator 'Temperature'. Values of the indicator are shown on the horizontal axis. Green points indicate states with no malfunction, red points indicate malfunctions (notation in German). The vertical axis serves better visualization and has no meaningful interpretation.

[9]Subjects are instructed to set their personal acceptable ranges to contain as many points without past malfunctions (green) and as few points with past malfunctions (red) as possible. This is not trivial, as there is no obvious dominant solution, just tendencies towards more or less efficient intervals.

The heuristic rule together with the individually defined acceptable ranges constitute the knowledge of subjects, necessary to take maintenance decisions. Additionally, subjects learn that they can utilize non-binding AI recommendations to aid their decisions. The accuracy of AI-recommendations can be 90% or 50%, depending on the success of the AI training[10]. In our experiment, the AI training is a separate process which takes place after subjects have set their acceptable ranges and before they are asked to take the maintenance decisions. The design of the AI training differs between the three experimental treatments, as required by our hypothesis.

In the zero involvement treatment, subjects are informed that the AI training is fully automated. They see a brief loading bar which indicates that the training is taking place. In the passive involvement treatment, subjects are told that the AI training takes place automatically based on their acceptable ranges. subjects are presented with 10 training situations. A training situation consists of a combination of three indicator values presented together with respective acceptable ranges[11] (see Figure 10). In each situation, the probability of a machine malfunction must be assessed using the heuristic rule. subjects are not allowed to assess the probability themselves. Instead, they observe how the heuristic rule is being automatically applied to each situation, based on their acceptable ranges. In the active involvement treatment, subjects are told that they actively train the AI with their knowledge. They are presented with the same 10 situations, but this time they can actively apply the heuristic rule and assess the malfunction probability by themselves. The design of the three treatments was chosen to mimic real-world involvement scenarios. Zero involvement serves as an equivalent to an AI training process which is carried out without domain expert participation, e.g., by data scientists or external providers. Passive involvement represents a situation in which domain experts share their knowledge about the system (in our design, in form of acceptable ranges), but do not actively apply this knowledge to the training data set. Instead, the knowledge would be processed and inserted into an AutoML system by an intermediary agent, e.g., a data scientist. Active involvement then allows domain experts to apply their knowledge directly to the training data set, without an intermediary. Based on the AutoML classification by Karmaker et al. [2020], the zero involvement scenario can be seen to represent ML solutions with up to the 3rd level of automation, passive and active involvement – the levels 4 and 5 respectively.

[10]The difference in the AI accuracy is introduced in order to encourage subjects to take the training seriously and to execute the necessary effort. Those who do not succeed during the training stage and receive a low-accuracy AI-advisor are excluded from the data analysis afterwards.

[11]From the training stage onward, acceptable ranges are displayed as gray bars while the historical data points are hidden for the purpose of conciseness.

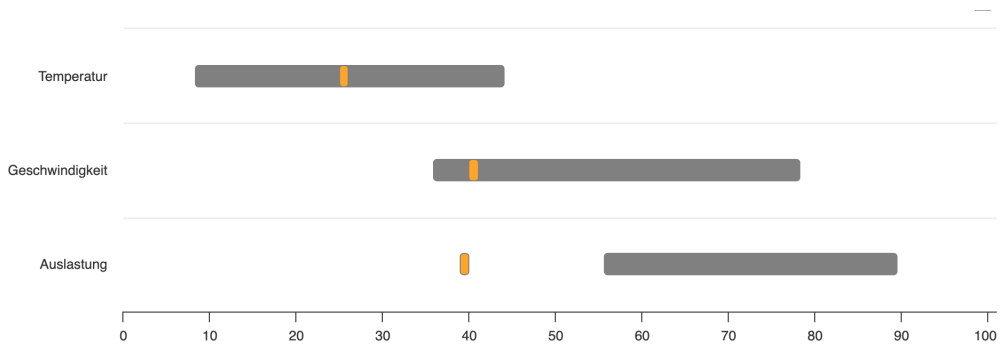


Figure 10: Example AI Training Situation. Individual acceptable ranges are displayed as gray bars, the indicator values are displayed as orange dots (notation in German).

If a malfunction has been classified as "very unlikely" or "unlikely" and no malfunction has occurred, the assessment is considered correct and otherwise incorrect. If a malfunction has been classified as "very likely" or "likely" and a malfunction has occurred, the assessment is considered correct and otherwise incorrect. In the zero- and passive involvement groups assessments are conducted automatically, based on the subjects' acceptable ranges and the heuristic rule. If at least seven training situations have been solved correctly, the AI-accuracy is 90%, otherwise 50%. In all three treatment groups, in the end of the AI training subjects are informed about the resulting AI-accuracy. The AI-accuracy is communicated to the subjects in both probability and frequency formats to ensure proper understanding [Gigerenzer and Hoffrage, 1995]. In the passive- and active involvement groups, subjects additionally learn, how many situations were solved correctly.

Finally, after the AI training is completed, the main part of the experiment starts. Subjects are asked to make maintenance decisions for a fictitious machine in 25 independent rounds. In every round, they are presented with a combination of three indicator values depicted together with the individual acceptable ranges (as in Figure 10). Each round can be represented as a decision-tree of three binary steps resulting in a total of eight well-defined decision paths (see Appendix B.3). In the first step, subjects are asked to assess the likelihood of a malfunction using the heuristic rule and to decide, whether a maintenance should be performed. In the second step, subjects are presented with the AI assessment of the malfunction likelihood as well as a recommendation on whether or not to perform a maintenance[12]. In the third step, subjects are asked to take their final maintenance decision, which will determine their payoff for the round. After a round is completed, subjects receive no immediate feedback on their performance in order to minimize learning effects as well as outcome-induced decision biases [Denrell and March, 2001]. Instead, subjects learn about their total performance and payoff once they have completed all 25 experimental rounds. Additionally, we purposefully do not communicate to subjects, how an average human subject would perform in a task at hand. Instead, we allow subjects

[12]The recommendations are were exogenously simulated beforehand, see Table 39 in Appendix 3.3.3, to ensure comparability between subjects

to develop a feeling of the own expertise, without explicitly quantifying it. This fact aims at mirroring an ambiguity of a real word situation, where people usually do not have access to an exact statistical information about their decision accuracy and the optimal strategy (to follow or not to follow the advice) is not straight-forward.

A payoff for a round is calculated as follows. If no maintenance is performed and no malfunction occurs, the machine can produce regularly, and subjects are granted 10 Taler[13]. Performing a maintenance reduces the round payoff to 5 Taler (a machine can produce only half of the regular amount) and eliminates the risk of a malfunction for that round. If no maintenance has been ordered and a malfunction occurs, subjects receive 0 Taler in this round (the machine had a zero production output due to the malfunction). Consequently, subjects must weigh a higher and risky payoff (when performing no maintenance) with a secure but smaller payoff (when performing a maintenance).

Subjects have been informed that the whole scenario is fictitious. The data used during the experimental scenario has been exogenously simulated beforehand from a random normal distribution. In total, 200 situations (combinations of three indicator values together with subsequent occurrence or absence of a malfunction) were generated. Out of these, 10 have been randomly assigned to the AI training situations, while 25 have been randomly assigned to the main part of the experiment. The AI-advisor has been simulated by a simple rule-based algorithm, which applied the heuristic rule while knowing the exact values of the optimal ranges. This procedure resulted in a high-accuracy AI-advisor that was 90% accurate and a low-accuracy AI-advisor with 50% accuracy. Each subject received either of the two depending on the outcome of the training stage.

3.2.2 Experimental Procedure

The experiment has been computerized using oTree [Chen et al., 2016] and conducted in December 2022 and April 2023 at the Business and Economic Research Laboratory (BaER-Lab; `www.baer-lab.org`.) at Paderborn University. In total, eight experimental sessions took place, three for each of the treatment groups. Subjects were recruited via the online recruiting system ORSEE [Greiner, 2015] and were only allowed to participate in one session. In the beginning of every session, subjects were randomly assigned to a computer workplace. Each workplace was placed in a cubicle to ensure that every subject could see only her own screen. Subjects were informed that communication and usage of mobile devices was prohibited during the experiment. Subjects received experimental instructions in printed form and had up to 15 minutes to read them carefully. Instructions contained all the information needed in the experiment and were tailored to respective treatments. Subjects were allowed to keep the instructions till the end of the experiment and to look them up at any time. The

[13]During the experiment, all amounts are denoted in the fictitious experimental currency "Taler" which is exchanged to Euro at €0.1 per Taler at the end of the experiment.

detailed instructions can be found in Appendix B.1.

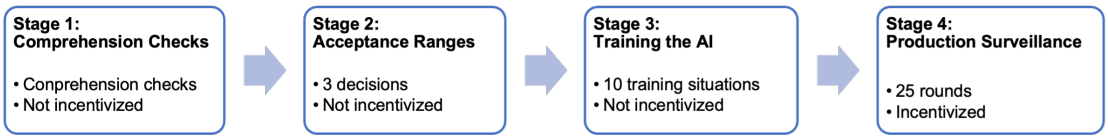


Figure 11: The Four Stages of the Experiment

The experimental process was logically divided into four stages (see Figure 11). In the first stage, subjects answered extensive comprehension questions to ensure a sufficient understanding of the experimental scenario and rules. Subjects could only advance to the second experimental stage once all questions had been answered correctly. In the second stage, subjects were asked to set their individual acceptable ranges for the three indicators. Beforehand, subject could practice setting acceptable ranges on an example distribution. During the setting, subjects were allowed to revise their ranges as many times as they wished until ultimately confirming their desired ranges and advancing to the third stage. In the third stage, the training of the AI took place. In the zero involvement group, subjects were simply informed about the result of the training without any details on its input or process. In the passive involvement group, the training happened automatically with no decision-making authority to subjects, but a possibility to observe the training process. Additionally, subjects knew that their acceptable ranges contributed to the AI training. In the active involvement group, subjects evaluated the training data themselves, applying their acceptable ranges and the heuristic rule, thereby directly contributing to the AI training. In stage four, subjects were asked to supervise the fictitious machine over 25 rounds. In each round they had to decide, whether a maintenance should or should not take place, going through the steps described in 3.2.1. All rounds of the fourth stage were independent of each other, i.e., a decision from one round did not affect other rounds, neither did an occurrence of a malfunction. The payoffs from all 25 rounds were accumulated and exchanged to Euro at a rate of €0.1 per Taler. Additionally, subjects received a show-up fee of €2.50. The total payoff was displayed after the completion of the stage four[14].

In the end of the experiment, subjects were asked to answer a questionnaire which contained multiple standardized questionnaires to account for the subjects' affinity for technology interaction [Franke et al., 2018], ex-post confidence in their decisions [Gino et al., 2012], and general self-efficacy [Beierlein et al., 2013], along with treatment manipulation checks and questions on the socio-economic background of the subjects, particularly age, gender, and study major[15]. Upon completion of the questionnaire, subjects were called by their cabin number to receive their payment. Subject were paid

[14]Stage four was the only stage to feature direct monetary incentives. However, at stages two and three subjects were informed that their decisions would have implications for their performance in stage four, thereby providing an indirect incentive to exert effort in the earlier stages as well.

[15]The full questionnaire can be obtained from Appendix B.2.

their earnings in cash and thanked for their participation before being released from the lab.

3.3 Results

The empirical analysis of our experimental observations was performed with Stata 17.0. First, we report demographics and descriptive statistics. Then, in the main analysis, we present the results of hypotheses testing and treatment effectiveness evaluation. We further conduct supplementary analysis on relevant questionnaire outcomes.

3.3.1 Demographic and Descriptive Statistics

In total, one-hundred-fifty-four (N=154) student subjects participated in the experiment, of which 51 were assigned to the zero involvement (T_Z), 51 to the passive involvement (T_P) and 52 to the active involvement (T_A) groups. Each session lasted approximately one hour, subjects earned €17.50 on average. One subject from T_Z failed to complete the comprehension questions and was therefore excluded from the data set (N=153). On average, subjects were 24 years old, their age varying between 18 and 65. Females constituted 64% of subjects (see Table 6). Fields of subjects' studies varied from engineering to arts, with economics (34%) and pedagogy (35%) being most common (see Table 40 in Appendix B.3).

Table 6: Demographic Statistics

	$\mathbf{T}_Z$	$\mathbf{T}_P$	$\mathbf{T}_A$	**Total**
Number of observations	50	51	52	153
Age				
Mean	24.4	24.5	23.1	24.0
Std. dev.	6.62	6.51	3.89	5.79
Gender (%)				
Female	72.0	62.7	57.8	64.1

Depending on the success of the AI training, subjects could receive an AI advisor with a high (90%) or low (50%) accuracy. The sole purpose of this AI accuracy manipulation was motivating subjects to exert effort during the AI training. For this reason, and to ensure the ceteris paribus condition between the treatments, subjects with a low AI accuracy ($n = 11$, therin 6 from T_P and 5 from T_A) are excluded from the following analysis. Observations are pooled at the subject-level, therefore the final data set includes $N = 142$ independent observations.

Table 7 summarizes the individual accuracy of initial assessments across the three treatments. On average, subjects' initial assessments were accurate in 52% ($\mathbf{T}_Z$), 58% ($\mathbf{T}_P$) and 57% ($\mathbf{T}_A$) of rounds. Although it seems that subjects from $\mathbf{T}_Z$ were slightly less accurate in their initial estimations, the differences in accuracy between the treatments are not statistically significant (Kruskal-Wallis H-Test

with ties: $\chi^2(2) = 3.041, p = 0.2186$). The standard deviation of initial accuracy is rather high at 17 p.p., 13 p.p. and 14 p.p. for $\mathbf{T}_Z$, $\mathbf{T}_P$ and $\mathbf{T}_A$ respectively. Analogously, the individual accuracy of initial assessments ranges between 0% and 88%. Logically, with the AI accuracy being at 90%, subjects with high initial accuracy had a lower probability of their initial estimation being different from the AI advice.

Table 7: Individual Accuracy of Initial Assessments, by Treatment

	$\mathbf{T}_Z$	$\mathbf{T}_P$	$\mathbf{T}_A$
Number of observations	50	45	47
Accuracy of initial assessments			
Mean (absolute)	13.1	14.5	14.3
Std. dev. (absolute)	4.3	3.4	3.4
Min	0	5	5
Median (absolute)	13	15	15
Max	22	21	20
Mean (relative)	0.52	0.58	0.57
Std. dev. (relative)	0.17	0.13	0.14
Median (relative)	0.52	0.60	0.60

Note: This table reports descriptive statistics on the (absolute and relative) frequency of individuals making assessments that are initially accurate over the 25 rounds across experimental groups.

After submitting their initial assessment, subjects received advice from the AI and could revise their maintenance decision. Subjects from the $\mathbf{T}_Z$ group revised their initial assessment in 11% of rounds and subjects from $\mathbf{T}_P$ and $\mathbf{T}_A$ groups in 8% of rounds (see Table 8). These percentages incorporate total revisions – both for rounds in which the initial assessment was equal or unequal to the AI advice. Subjects from $\mathbf{T}_Z$ appear to have revised their assessments slightly more often, however the differences between the treatment groups are not statistically significant (Kruskal-Wallis H-Test with ties: $\chi^2(2) = 3.907, p = 0.1418$). Standard deviations were 9 p.p., 8 p.p. and 8 p.p. for $\mathbf{T}_Z$, $\mathbf{T}_P$ and $\mathbf{T}_A$ respectively. Notably, we observe both subjects who never revised their decision as well as those who revised their decision in nearly 50% of rounds. Among all revisions, 70% to 82% eventually led to a better outcome for the subjects (see Table 41 in Appendix B.3).

Table 8: Revision of Initial Decisions, by Treatment

	$\mathbf{T}_Z$	$\mathbf{T}_P$	$\mathbf{T}_A$
Number of observations	50	45	47
Revision of initial assessments			
Mean (absolute)	2.7	2.0	2.1
Std. dev. (absolute)	2.3	2.0	2.0
Min	0	0	0
Median (absolute)	3	2	2
Max	8	12	9
Mean (relative)	0.11	0.08	0.08
Std. dev. (relative)	0.12	0.08	0.08
Median (relative)	0.12	0.08	0.08

Note: This table reports the (absolute and relative) frequencies of subjects revising their initial decision after receiving algorithmic advice across experimental groups.

3.3.2 Main Analysis

Following our research hypotheses, we are particularly interested in situations, in which subject´s initial assessment differed from the AI's advice. For such situations, if the revised decision was equal to the AI advice, we count it as AI adherence. Per subject, we then calculate an "AI adherence rate" – the proportion of AI adherence among situations where one's initial assessment differs from the AI advice. Table 9 summarizes the corresponding empirical results for the three treatment groups. On average, subjects' initial assessment contradicted the AI's advice in 5 ($\mathbf{T}_Z$) or 4 ($\mathbf{T}_P$, $\mathbf{T}_A$) out of 25 experimental rounds, at standard deviations of 4 ($\mathbf{T}_Z$) and 3 rounds ($\mathbf{T}_P$, $\mathbf{T}_A$) respectively. While subjects' initial assessments in $\mathbf{T}_Z$ deviated from the advice slightly more often, the difference is not significant (Kruskal-Wallis H-Test with ties: $\chi^2(2) = 1.89, p = 0.3887$). Consequently, subject had slightly more opportunities to follow the AI's advice in $\mathbf{T}_Z$. However, this On average, the AI adherence rate was 48% in the zero involvement, 51% in the passive involvement and 47% in the active involvement group, with rather high standard deviations at 35p.p., 37p.p. and 36p.p. respectively. Notably, in all three groups there are subjects who either always or never follow the AI advice. It is also worth mentioning that subjects who revised their assessment following contradictory AI advice benefited from it in 80% to 86% of instances (see Table 41 in Appendix B.3). However, the efficiency of this strategy was not entirely apparent to the subjects, as they were unable to directly assess their own accuracy during the experiment due to the lack of feedback provision on previous rounds.

Table 9: AI Adherence Rate, by Treatment

	$\mathbf{T}_Z$	$\mathbf{T}_P$	$\mathbf{T}_A$
Number of observations	50	45	47
Initial assessment ≠ AI advice			
Mean (absolute)	5.1	3.8	4.1
Std. dev. (absolute)	4.0	2.6	2.8
Min	0	0	0
Median (absolute)	4	4	4
Max	17	12	12
Mean (relative)	0.206	0.154	0.163
Std. dev. (relative)	0.161	0.103	0.111
Number of observations	46	43	43
AI advice adherence rate			
Mean	0.481	0.508	0.466
Std. dev.	0.348	0.374	0.356
Median	0.4	0.5	0.5
Min	0	0	0
Max	1	1	1

Note: This table reports the instances in which subjects' initial assessment was unequal to the AI advice, as well as the AI advice adherence rate. Ten subjects whose initial assessment has been equal to the AI advice in all 25 rounds have been excluded from the calculation of the AI adherence rate.

Figure 12 depicts the frequency distributions of AI adherence rates for the three treatment groups. Although each group's distribution is heterogeneous within, the patterns between the groups appear to be similar. The distribution of the AI adherence rates between 0 and 1 takes a "W"-like shape, indicating that the boundary solutions – always or never following the AI advice – are the prevailing strategies in every group. This meets findings from related literature of individuals tending to be either averse or appreciative of algorithmic advice on a given task [Jussupow et al., 2020].

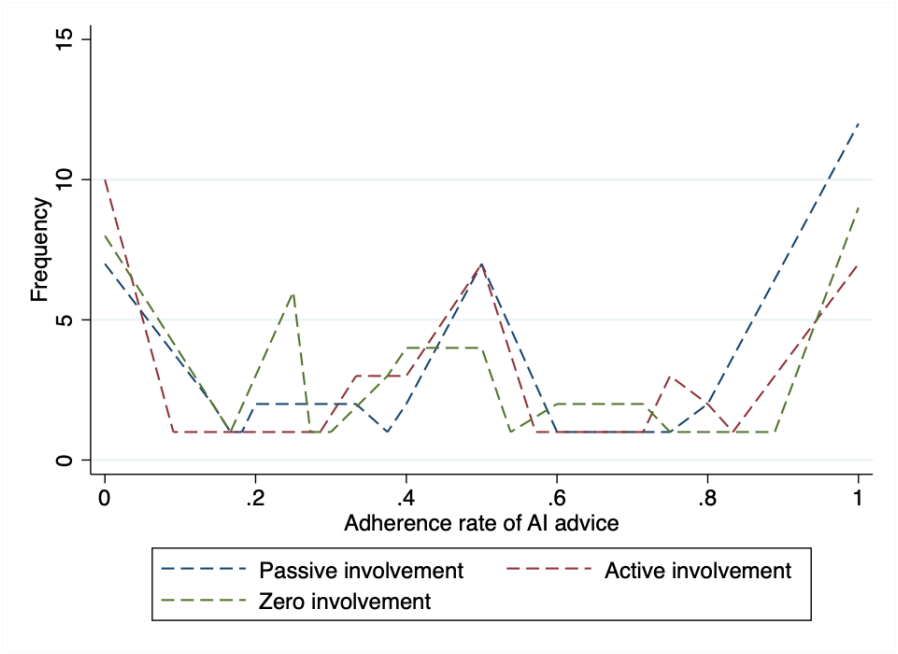


Figure 12: Distribution of AI Adherence Rate, by Treatment

To investigate our first research hypothesis (Hypothesis 1a), we conduct a Kruskal-Wallis H-Test with ties. We find no statistically significant difference in the AI adherence rate between the three treatment groups ($\chi^2(2) = 0.299, p = 0.8610$). Pairwise comparisons of the groups yield the same results (Mann-Whitney U-Test, one-sided, each subject one independent observation: $\mathbf{T}_P$ vs. $\mathbf{T}_A$: $z = 0.529, p = 0.3000$; $\mathbf{T}_A$ vs. $\mathbf{T}_Z$: $z = 0.070, p = 0.4729$; $\mathbf{T}_P$ vs. $\mathbf{T}_Z$: $z = 0.423, p = 0.3375$). Additionally, we run pairwise two-sample Kolmogorov–Smirnov tests for equality of distribution functions. The tests show that distributions of AI adherence rates (between 0 and 1) do not significantly differ between groups ($\mathbf{T}_P$ vs. $\mathbf{T}_A$: $D = 0.1163, p = 0.938$; $\mathbf{T}_A$ vs. $\mathbf{T}_Z$: $D = 0.1067, p = 0.925$; $\mathbf{T}_P$ vs. $\mathbf{T}_Z$: $D = 0.1097, p = 0.908$)), further supporting the results from the two previous tests as well as the impression gained from Figure 12.

Consequently, we conclude that the hypotheses Hypothesis 1a is not supported by our empirical data. Since our second hypothesis (Hypothesis 1b) requires at least one significant difference between any of the treatment groups, it is therefore also not supported. To shed light on possible causes behind the rejection of the main hypotheses, we evaluate the effectiveness of our experimental design. For our experiment, it is crucial that subjects perceive themselves as experts within the experimental scenario, independently of the treatment. Further, the treatment manipulations should lead to differences in the subjects' perception of the AI. Table 10 summarizes subjects' perception of the experimental scenario. The results demonstrate that subjects from all treatment groups understood the scenario and its rules well, were able to put themselves into the role of an engineer, and were satisfied with their own performance. The scores for all items appear comparably high, at 3.8 and upwards on a 5-point scale.

The only exception is the perceived demand of the experimental tasks, with scores between 2.6 and 2.8, which shows that the task was neither perceived as very easy or very difficult by the subjects.

Table 10: Summary Statistics and Between-Group Comparison of Questionnaire Items on Experiment Perception, by Treatment

	$\mathbf{T}_Z$	$\mathbf{T}_P$	$\mathbf{T}_A$	K-Wallis
Items on experiment perception				
Maintenance scenario was easily understandable	4.64	4.53	4.66	0.520
	(0.63)	(0.63)	(0.52)	
Able to imagine oneself as production manager	3.9	3.82	4.09	0.632
	(1.15)	(1.17)	(0.95)	
Sufficient information and examples provided	4.6	4.51	4.51	0.479
	(0.70)	(0.63)	(0.66)	
Understood experimental explanations	4.68	4.58	4.77	0.298
	(0.68)	(0.62)	(0.43)	
Understood experimental context	4.72	4.67	4.83	0.267
	(0.50)	(0.52)	(0.38)	
Certainty about experimental requirements	4.78	4.71	4.77	0.824
	(0.46)	(0.59)	(0.52)	
Understanding of conditions for success	4.42	4.38	4.49	0.709
	(0.76)	(0.69)	(0.66)	
Perceived task demand	2.58	2.78	2.55	0.456
	(1.01)	(0.95)	(0.88)	
Satisfaction with own performance	4.3	4.42	4.06	0.192
	(0.65)	(0.54)	(0.92)	
Expectation of other subjects performing well	3.84	3.93	4.04	0.472
	(0.84)	(0.75)	(0.69)	
Number of observations	50	45	47	

Note: This table reports summary statistics of questionnaire items on experiment perception (measured on a 5-point scale). Standard deviations reported in parenthesis. "K-Wallis" reports p-values for Kruskal-Wallis H-Tests with ties between experimental groups.

Further, we evaluate questionnaire items relevant for our treatment manipulations: perceived influence on AI training, perceived contribution to AI advice quality, perceived understanding of AI functionality, perceived understanding of AI advice generation and perceived quality of AI advice. Tables 11 and 12 summarizes the corresponding statistics[16]. The first four items each demonstrate a pattern of subjects' consent to questionnaire statements gradually increasing from zero to active involvement treatments. In other words, subjects perceived influence on AI and understanding of AI increases with the degree of their involvement, with the differences between $\mathbf{T}_Z$ and $\mathbf{T}_A$ being highly statistically significant for all items, while the differences between $\mathbf{T}_P$ and $\mathbf{T}_A$ are significant for half of them. Overall, we conclude that our treatment manipulations were successful in the sense of different grades of subjects' involvement in the AI training being reflected in subjects' perception of their understanding of and influence on the AI. However, these gradations in perception are not reflected in subjects' actual behavior, as the rejection of our main hypotheses demonstrates.

Generally, the scores for subjects' perceived influence on AI and understanding of AI appear rather

[16] A full list of items as well as between-treatment comparisons can be found in Table 42 Appendix B.3.

high – even in the zero-involvement group, subjects perceive a certain degree of influence and understanding with scores of around 3 points on a 5-points scale[17]. While subjects' perceived quality of the AI advice also appears high at 3.6 points and upwards across all treatment groups, there are, however, no significant differences between groups in this regard.

Table 11: Questionnaire Items on AI Perception, by Treatment

	$\mathbf{T}_Z$	$\mathbf{T}_P$	$\mathbf{T}_A$	K-Wallis
Items on AI perception				
Perceived influence on AI training	3.02	3.58	4.09	0.0001
	(0.91)	(1.12)	(0.80)	
Perceived contribution to AI advice quality	3.34	3.87	4.11	0.0013
	(1.10)	(1.06)	(0.91)	
Perceived understanding of AI functionality	3.08	3.38	3.91	0.0024
	(1.21)	(1.23)	(1.02)	
Perceived understanding of how AI generates advice	2.9	3.42	3.72	0.0020
	(1.15)	(1.16)	(1.16)	
Perceived quality of AI advice	3.68	3.76	3.60	0.4874
	(0.79)	(1.05)	(0.95)	
Number of observations	50	45	47	

Note: This table reports means for AI perception items (5-point scale) with standard deviations in parenthesis. "K-Wallis" reports p-values for Kruskal-Wallis H-Tests with ties.

Table 12: Pairwise Between-Group Comparison of Questionnaire Items on AI Perception

	Mann-Whitney U-test (p-value)		
Pairwise comparison	$\mathbf{T}_Z$ **vs.** $\mathbf{T}_P$	$\mathbf{T}_Z$ **vs.** $\mathbf{T}_A$	$\mathbf{T}_P$ **vs.** $\mathbf{T}_A$
Perceived influence on AI training	0.0063	0.0000	0.0270
Perceived contribution to AI advice quality	0.0186	0.0003	0.2908
Perceived understanding of AI functionality	0.1995	0.0005	0.0362
Perceived understanding of how AI generates advice	0.0284	0.0005	0.1645
Perceived quality of AI advice	0.3456	0.7550	0.2726

Note: This table reports p-values for pairwise two-sample between-group Mann-Whitney U-Tests between experimental groups.

3.3.3 Supplementary Analysis

To control for factors that potentially influence individuals' AI adherence, we measured subjects' affinity for technology interaction (ATI), general self-efficacy (ASKU) and ex-post decision confidence (DC), using standardized questionnaire metrics. Table 13 summarizes the corresponding statistics and between-group comparisons. The average ATI scores show moderate technological affinity among subjects, ranging between 3.5 and 4 on a 6-point scale. The average ASKU score resides in the upper quartile of the 5-point scale for all groups, indicating that subjects generally perceive themselves capable of dealing with problems and accomplishing goals. Similarly, subjects appear quite confident

[17]Consistent with this observation, subjects assign rather low scores to the question, whether AI advice accuracy is a gamble, see Table 42 in B.3.

about their decisions made during the experiment, with the average ex-post confidence for all three groups exceeding 5 on a 7-point scale.

Table 13: Comparison of Standardized Questionnaire Controls, by Treatment

	$\mathbf{T}_Z$	$\mathbf{T}_P$	$\mathbf{T}_A$	K-Wallis
Affinity to Technology Index (ATI)	3.46	3.55	4.01	0.0460
	(0.89)	(1.21)	(1.13)	
Self-efficacy (ASKU)	3.77	3.78	3.89	0.2950
	(0.54)	(0.54)	(0.54)	
Ex-post decision confidence (DC)	5.19	5.27	5.08	0.7657
	(0.85)	(0.82)	(1.11)	
Number of observations	50	45	47	

Note: Standard deviations are reported in parenthesis. ATI is measured on a 6-point scale, DC is measured on a 7-point scale, ASKU is measured on a 5-point scale. "K-Wallis" reports p-values for Kruskal-Wallis H-Tests with ties between experimental groups.

We do not find significant treatment differences for general self-efficacy (Kruskal-Wallis H-Test with ties: $\chi^2(2) = 2.442, p = 0.2950$) and ex-post confidence in decisions (Kruskal-Wallis H-Test with ties: $\chi^2(2) = 0.534, p = 0.7657$). Notably, the data reveals a weak yet statistically significant difference in the affinity for technology interaction between the treatment groups (Kruskal-Wallis H-Test with ties: $\chi^2(2) = 6.156, p = 0.0460$), which can be attributed to the differences between $\mathbf{T}_Z$ and $\mathbf{T}_A$ (Mann-Whitney U-Test, two-sided: $z = 2.514, p = 0.0115$)[18]. However, this difference must be interpreted with caution, as affinity for technology interaction was measured after the experiment. As $\mathbf{T}_Z$ and $\mathbf{T}_A$ present polar opposites in terms of technology interaction quantity, this may have factored into subjects' assessment of their affinity for technology interaction.

Additionally, subjects in all groups considered the AI training a time well spent (3.7-4 points, on average), indicated a moderate enjoyment (3.3-3.7 points), and liked sharing knowledge with AI (3.9-4 points) (see Table 42 in Appendix B.3). The scores for opposite items – unwillingness and discomfort in sharing knowledge with AI – are consistently low and do not exceed 2 points.

3.4 Discussion and Conclusion

Today's world witnesses a rapidly increasing relevance of artificial intelligence with exponentially growing number of its application domains. However, development of ML models requires programming and data science competence which is scarce on the labor market. Automated machine learning is supposed to make AI solutions more accessible to organizations without sufficient data science expertise. Domain experts play a crucial role in making this vision a reality, acting as knowledge contributors. We argue that besides providing the domain knowledge, involvement of domain experts has an ad-

[18]For the full set of pairwise Mann-Whitney U-Tests see Table 43 in Appendix B.3

ditional value, since they are in many cases also the end users of AutoML solutions. By involving domain experts in the model training, one enables them to execute influence over the resulting model and to enhance their understanding of its functionality. Drawing from pre-existing literature, these two factors in their turn may strengthen domain experts trust in the AutoML model and raise their adherence to its recommendations. The adherence is crucial for organizations if they want to turn the high accuracy of AI advisor systems into real efficiency gains.

Our study aims to investigate the role of domain expert involvement during the AI training on their subsequent AI advice adherence. We hypothesize that a higher grade of involvement leads to a higher advice adherence. In an incentivized laboratory experiment, we construct a predictive maintenance scenario in which subjects act as domain experts and are confronted with an AI advisor. We compare three groups: subjects who actively participate in the AI training, subjects who passively witness the AI training, and subjects who neither participate nor observe the AI training. We measure the individual adherence rate for each subject, which reflects the ratio between the number of decisions to follow an AI recommendation and the number of situations, in which the subjects' initial estimation is different from the AI advice. In our main analysis, we compare the adherence rates between the treatment groups.

We find that our treatment manipulations had the intended effect on the subjects' perception – the higher the degree of involvement in the AI training, the higher the perception of own influence on the AI model. Additionally, subjects' perception of their own understanding of the AI functionality also grows with a higher grade of involvement. According to the related literature, both higher perceived influence and higher perceived understanding should lead to a higher perceived quality of the AI advice and eventually to a higher advice adherence [Dietvorst et al., 2015; Jussupow et al., 2020; Mahmud et al., 2022; Yeomans et al., 2019]. However, we observe subject who actively participate in the AI training follow the AI advice similarly often as subjects from passive and zero involvement groups. Therefore, we have to reject both initial hypotheses as adherence behavior does neither vary between groups nor grow with the increasing grade of involvement. These results reflect prior findings gaps between stated trust in and actual reliance on algorithmic systems [Schmitt et al., 2021].

We conclude that involvement of domain experts in the AI training enhances the feeling of control over the resulting model and the perceived understanding of the AI's functionality, as it helps them influence the model, fit it to their domain context, and peer inside the black box, although apparently no behavioral change results form this. Both these insights are valuable in the context of the human-centered AI and AI explainability [Bingley et al., 2023; A. and R., 2023].

One possible explanation for the fact that the positive effects of involvement on AI perception do not translate into behavior may be that these effects are overruled by the strong feeling of expertise among

subjects, which did not depend on treatment. Generally, across all three involvement conditions, we find adherence to AI advice to be rather low, at roughly 50%, despite subjects being informed about the AI's accuracy of 90%. Simultaneously, subjects report to have understood the scenario well and to know exactly, what they should do in order to be successful in the tasks. These insights, together with the high reported ex-post decision confidence, support the intended of subjects assuming to be experts within the experimental scenario. Subjects feel equally (and strongly) confident in their choices and might disregard AI advice equally often because of this. Experts have been repeatedly shown to behave differently in algorithm interactions than lay people and in particular to demonstrate higher algorithm aversion [Logg et al., 2019; Mahmud et al., 2022]. Our findings reinforce these results, although they should be interpreted with due caution, since our experiment featured subjects with expertise in a given task rather than actual expert practitioners.

We further observe a rather heterogeneous distribution of algorithm adherence behavior within each group with tendencies towards the extremes, i.e., individuals either acting very averse or very appreciative towards AI-advice with only a minority adapting their decisions situation-dependent. This matches the state of the existing literature on algorithm adherence behavior [Jussupow et al., 2020], with these within-group differences appearing to be consistent between groups. Also, other individual factors, like affinity for technology interaction or general self-efficacy do not vary significantly between the treatment groups. Interestingly, subjects report strong positive perceptions towards the AI training – they consider it to be a well-spent time, share their knowledge willingly and even enjoy the processes. These scores are equally high across all three treatment groups and go along with the "IKEA-effect", repeatedly found in the previous literature [Norton et al., 2012].

Future research may investigate whether feedback on own accuracy could relativize subjects' perception about their own expertise and make them more accessible to the treatment manipulations. However, learning effects would have to be taken into account in this case. Further, the number of single observations, in which the AI adherence rate could be calculated, has been rather small. By providing subjects with expertise through the rule-of-thumb for production indicator assessment, we enable them to make accurate maintenance decisions. This high initial accuracy reduces the number of instances in which the own initial assessment differs from the AI advice (4 to 5 out of 25 rounds, on average) – mirroring the real world where experts would indeed often arrive at the same assessment as the AI, even though the AI advisor would likely prove more accurate over time. However, from an experimental standpoint, this substantially reduces the number of informative observations. Future iterations of the experiment could address this issue by increasing the total number of rounds to generate more instances in which AI adherence can be meaningfully measured. Increasing task uncertainty would be another potential direction to go in. Also, other forms of expert involvement and knowledge

sharing can be considered – an interview, a questionnaire, etc. Finally, future research can test the direct influence of controllability and understandability on advice adherence, for example by explicitly allowing involvement to change the model output [Dietvorst et al., 2018; Gubaydullina et al., 2022] or explaining the functionality of AutoML to domain experts [Yeomans et al., 2019].

As we introduced a novel experimental design through which behavioral effects of user involvement in AI-supported predictive maintenance decisions can be studied, researchers may build upon our design to investigate a variety of further research questions regarding human perception of algorithmic entities, particularly in industry or production-centric contexts. Overall, future research should continue to pursuit the goal of fostering algorithm adherence to mitigate economic inefficiencies induced by averse (and overly appreciative) individual behavior.

In any case, our results carry at least two positive signals for practitioners: Through end-user involvement, they can improve their (perceived) understanding of the AutoML system. Additionally, users appear open to sharing knowledge with AI.

Chapter 4:

"Behavioral Economics for Human-in-the-loop Control Systems Design: Overconfidence and the Hot Hand Fallacy"

Behavioral Economics for Human-in-the-loop Control Systems Design: Overconfidence and the Hot Hand Fallacy

Marius Protte[*,†], René Fahr[*], Daniel E. Quevedo[#]

Abstract

Successful design of human-in-the-loop control systems requires appropriate models for human decision-makers. Whilst most paradigms adopted in the control systems literature hide the (limited) decision capability of humans, in behavioral economics individual decision-making and optimization processes are well-known to be affected by perceptual and behavioral biases. Our goal is to enrich control engineering with some insights from behavioral economics research through exposing such biases in control-relevant settings. This paper addresses the following two key questions: 1) How do behavioral biases affect decision-making? 2) What is the role played by feedback in human-in-the-loop control systems? Our experimental framework shows how individuals behave when faced with the task of piloting an UAV under risk and uncertainty, paralleling a real-world decision-making scenario. Our findings support the notion of humans in cyber-physical Systems underlying behavioral biases regardless of – or even because of – receiving immediate outcome feedback. We observe substantial shares of drone controllers to act inefficiently through either flying excessively (overconfident) or overly conservatively (underconfident). Furthermore, we observe human-controllers to self-servingly misinterpret random sequences through being subject to a "hot hand fallacy". We advise control engineers to mind the human component in order not to compromise technological accomplishments through human issues.

JEL Classification: D81, D82, D91, C91

Keywords: Bounded rationality; Markov decision-maker; drone piloting; overconfidence; hot hand fallacy

[*]Paderborn University, Heinz-Nixdorf-Institute, Fürstenallee 11, 33102 Paderborn

[#]School of Electrical Engineering & Robotics, Queensland University of Technology, QLD 4000, Brisbane, Australia

[†]Corresponding author, marius.protte@upb.de

This article was published in *IEEE Control Systems Magazine*, DOI: https://doi.org/10.1109/MCS.2020.3019723. The present version deviates from the published version in minor textual and organizational changes. This research was funded by the Deutsche Forschungsgemeinschaft within the "SFB 901: On-The-Fly (OTF) Computing – Individualised IT-Services in Dynamic Markets" program (160364472).

4.1 Introduction

Our century is bringing interesting challenges and opportunities that derive from the way that digital technology is shaping how we live as individuals and as a society. A key feature of a number of engineered systems is that they interact with humans. Rather than solely affecting humans, often people make decisions that have an effect on the engineered system. As an example, when driving our cars we often decide to take a route which differs from that suggested by the navigation system. This information is fed back to the service provider and henceforth used when making route suggestions to other users. The analysis and design of such *Cyber-physical Human Systems* (CPHSs) would benefit from an understanding of how humans behave. However, given their immense complexity, it is by no means clear how to formulate appropriate models for human decision-makers, especially when operating in closed loop systems.

CPHSs fit into the general context of human-machine systems and pose multidisciplinary challenges that have been tackled in a number of domains. For example, interesting surveys on signal processing approaches include Vempaty et al. [2018] and Narayanan and Georgiou [2013], whereas Schirner et al. [2013] adopts a computer science viewpoint. More focused on specific applications, the articles in the special issue Tanelli et al. [2017] as well as Dressler [2018] study the interaction between humans and vehicles. Kolling et al. [2016] surveys systems comprised of humans and robot swarms and Young and Peschel [2020] reviews telemanipulation by small unmanned systems. Also within the control systems community, significant advances have been made [see, e.g., Hokayem and Spong, 2006; van Overloop et al., 2015; Ercan et al., 2017; Inoue and Gupta, 2019]. Here it is common to model the effect of humans as a limited actuation resource or unknown deterministic dynamics to be identified. This opens the door to use various robust control and game theoretic methods to design closed loop systems. However, most paradigms adopted in the systems control literature so far hide the decision capability of humans and limitations to their cognitive and computational capabilities. Notable exceptions to the literature include Lam and Sastry [2014] and Feng et al. [2016] where human decision-making is characterized via a Partially Observed Markov Decision Process (MDP) [Puterman, 1994] with states representing basic physical human aspects, such as operator fatigue or proficiency. In particular, Feng et al. [2016] investigates a stochastic two player game, resulting from the interaction between a human operator and an unmanned aerial vehicle (UAV). Here the UAV is allowed to react to the decisions made by the operator, to compensate for potential non-cooperative behavior. For a CPHS with multiple human decision-makers, Albaba and Yildiz [2019] models limitations in their decision-making capabilities using "level-k reasoning" and associated game theory [Camerer et al., 2004]: Each agent optimizes a cost function using only reduced knowledge of the other agents' policies.

In the present work we investigate human decision-making from a behavioral economics perspective. We pose the question how behavioral biases affect decision-making, since we expect behavioral influences to potentially hamper technological optimization efforts. More generally, we aim to direct attention to the human as a poorly observed uncertainty source in CPHS. We thus consider humans as being "bounded rational", i.e., they are strategic thinkers who want to maximize their benefits but, at the same time, have only limited cognitive and computational capabilities. We distinguish between closed loop scenarios, where the decision-maker receives feedback about the success of its actions, and open loop situations, where no such feedback information is available. Here, we pay special attention to the role played by feedback on behavioral biases in form of overconfidence [Schaefer et al., 2004] and the so-called hot hand fallacy [Gilovich et al., 1985]. The latter effect may arise when a decision-maker becomes aware of her past success and overestimates future success probabilities. We will next introduce some prominent findings from behavioral economics that can be regarded as influential for CPHSs. These will be subsequently applied to an experimental human-in-the-loop framework inspired by the UAV piloting scenario of Feng et al. [2016]. Throughout our presentation we will highlight the value of giving greater consideration to human behavior and behavioral biases in control engineering. The central outcome of our study is that frequent feedback to human decision-makers may lead to sub-optimal results. This stands in contrast to situations wherein computers carry out optimizations and more information is always beneficial [Bar-Shalom and Tse, 1974; Puterman, 1994; Bertsekas, 2005]. As we shall see, when humans act as decision-makers, then the situation is more ambiguous.

4.2 Contributions from Behavioral Economics

Behavioral economics is a comparably young research field that is concerned with the overarching question of how humans actually do behave in economic decisions, in contrast to how they are prescribed to behave by economic theory [Kahneman, 2003]. Therefore, traditional economic assumptions and models are revised and enriched by insights from psychology. This serves to improve the understanding and predictability of human behavior, especially regarding recurring errors and biases in decision-making [Camerer and Loewenstein, 2003]. The perception of humans thereby shifts from the assumption of a rationally optimizing Markov decision-maker towards regarding humans as rather bounded rational agents. Humans constantly use cognitive mechanisms of simplification, namely decision heuristics, in order to process information and make decisions under uncertainty [Kahneman, 2003; Mousavi and Gigerenzer, 2014]. Such heuristics are used subconsciously due to individuals not managing to adequately process the complexity of a decision problem or to take all relevant information into account [Camerer and Loewenstein, 2003]. In fact, when making judgments or estimations of events, frequencies or probabilities under uncertainty, individuals do not always obey Bayesian rules and statistical logic,

as they are meant to do in models assuming perfectly rational agents and Markov decision-makers. Such heuristic simplifications sometimes yield reasonable judgments, but may also lead to severe and systematic errors [Kahneman and Tversky, 1973; Tversky and Kahneman, 1974; Todd and Gigerenzer, 2003]. Interestingly, even experienced researchers and professionals often underlie the same judgment heuristics and biases as laypersons do [Tversky and Kahneman, 1974].

4.2.1 Overconfidence

Overconfidence is well-known to be one of multiple behavioral stylized facts influencing and thereby biasing human decision-making. Overconfidence can be defined as a general miscalibration in beliefs [Lichtenstein and Fischhoff, 1977], more specifically, the discrepancy between confidence and accuracy [Schaefer et al., 2004]. In being overconfident, people overestimate their own capabilities, their knowledge or the general favorability of future prospects [Barber and Odean, 2001]. DeBondt and Thaler [1995] go as far as labeling overconfidence the "most robust finding in the psychology of judgment". Early contributions to research on overconfidence found that people systematically tended to be unrealistically optimistic about their future, in judging themselves to be more likely to experience a variety of positive events, and to be less likely to experience negative events, compared to others. This pattern has been traced back mainly to the degree of desirability influencing the perceived probability of such events [Weinstein, 1980]. In the original study by Svenson [1981], over 80% of the subjects regarded themselves as more skillful and less risky car drivers than the average driver. Moreover, one half of the subjects estimated themselves to be among the top 20% of the sample, vividly illustrating overconfidence as the discrepancy between confidence and accuracy (belief and reality) in doing so.

Overconfidence can express itself, and consequently has been studied, in multiple ways, like "better-than-average" beliefs, [see, e.g., Hoelzl and Rustichini, 2005], or "overprecision", [see, e.g., Radzevick and Moore, 2011]. The facet of overconfidence most fitting for the purpose of our current work appears to be overestimation, described as self-servingly overestimating the likelihood of desirable outcomes, supposedly fueled by wishful thinking for such desirable outcomes to occur[19] [Moore and Schatz, 2017; Mayraz, 2011]. While evidence on overconfidence generally does not depend on participants' familiarity with a task, it is influenced by task difficulty, with more challenging tasks tending to elicit greater overestimation than easier ones [Lichtenstein and Fischhoff, 1977; Moore and Healy, 2008]. Experimental approaches to studying overconfidence were requested by Hoelzl and Rustichini [2005], as most research in this field merely relies on verbal statements or subjective estimations, rather than monitoring human decision-behavior. Our current analysis makes a contribution to this call.

However, overconfidence is connected to a variety of other behavioral phenomena, which will be

[19] Overconfidence is therefore regarded as closely related to self-deception [see, e.g., van Hippel and Trivers, 2011].

reviewed in the sequel.

4.2.2 Underestimation of Systematic Risk

Analogous to overestimating favorable outcomes, overconfidence also implies underestimating risks, or rather the variance of risky processes, indicating a too narrow distribution in one's subjective probability beliefs [Ben-David et al., 2013]. Often, humans assess probabilities incorrectly or rather draw incorrect conclusions from them through not adhering to Bayesian rules or neglecting base-rates [Kahneman and Tversky, 1973; Sedlmeier and Gigerenzer, 2001]. In doing so, smaller probabilities are usually overestimated while larger probabilities are underestimated [Kahneman and Tversky, 1979; Slovic et al., 1982]. Vice versa, decision-makers are also found to underestimate the overall likelihood of the occurrence of risks with small single probabilities of occurrence in some instances. Especially risks of disqualification [Abbink et al., 2002] through low-probability events generate less concern than their probability warrants on average [Weber, 2006]. A cognitive process behind this underestimation of especially very low probability risks, like being involved in car crashes or natural disasters, can be characterized as subconsciously approximating low probabilities with zero in order to not having to mentally deal with them anymore [Slovic et al., 1978; Hogarth and Kunreuther, 1989].

Consequences of this issue can be observed in various areas of economic and social decision-making. Overconfident individuals, who underestimate health, financial or driving risks, are more likely to have insufficiently low insurance coverage against such risks [Sandroni and Squintani, 2013]. Furthermore, employees are observed to underestimate the risk of their own company's stock and to be overly optimistic about its future performance. They tend to include such stocks too heavily into their retirement plans, despite the respective stock being riskier than the overall market, as a consequence of excessive extrapolation of positive past performance [Benartzi, 2001]. Money being at stake does neither change overconfident behavior, nor does it lead to better estimation results concerning abilities, probabilities or risk. Experimental evidence was obtained on overconfidence leading to overly rushed market entries, that often precede business failures, due to entrepreneurs overestimating their relative chances to succeed Camerer and Lovallo [1999]. Overconfident top-managers were further shown to underestimate the volatility of cash flows of S&P 500 companies, resulting in erroneous investment and financing decisions, in a large-scale survey over 10 years [Ben-David et al., 2013]. Especially managers competing for leadership positions display overconfidence by tending to take on riskier projects due to underestimating their risks [Goel and Thakor, 2008]. Overconfident CEOs also overestimate the positive impact of their leadership and their ability to successfully complete a merger to generate future company value. These CEOs are found to execute value-destroying mergers, thereby exposing the affected companies to high financial or even existential risks [Malmendier and Tate, 2005]. Multiple experimental stud-

ies found that overconfident investors trade more excessively than others, overestimate the impact of little information they have [Odean, 1999; Barber and Odean, 2000], and do not adjust their trading volume to new negative information [Trinugroho and Sembel, 2011]. These findings, which are inconsistent with the Markov assumption and Bayesian updating, contradict theoretical expectations of rationally optimizing economic agents and re-emphasize the notion of individuals as bounded rational decision-makers [Kahneman, 2003].

4.2.3 Attribution Theory

Another mechanism likely contributing to overconfident decision-making is a biased perception of causality in a self-serving way. Accordingly, individuals express the tendency to attribute positive past outcomes to themselves and their abilities, while blaming negative outcomes on external circumstances [Frieze and Weiner, 1971] – successes are internalized, failures are externalized [Langer, 1975]. In an early experiment, teachers attributed learning successes of their pupils to their teaching skills, while blaming bad learning performances on the pupils themselves [Johnson et al., 1964]. This self-serving attribution bias was identified as the major driver of the aforementioned CEO overconfidence leading to risky mergers [Billett and Qian, 2008]. In a study by Gervais and Odean [2001], overconfidence evolves when traders receive feedback about their ability through experience in a multi-period market model and overweigh the role of their ability on prior success. Overconfident investors subsequently tend to become even more overconfident in their future investments. Meanwhile, financial losses are rather attributed to environmental circumstances, such as unfavorable macroeconomic developments or simply bad luck, with the investor's degree of overconfidence remaining constant [Hilary and Menzly, 2006; Daniel and Hirshleifer, 2015]. Also, professionals were found to be as likely as lay people to express overconfidence in making economic decisions as well as in re-evaluating the quality of their own previous decision in hindsight [Törngren and Montgomery, 2004].

People overestimate their own capabilities as well as their control over future events that are actually determined by external factors, especially chance. This was illustrated in an experimental study in which participants were either given a lottery ticket (control group) or allowed to select a lottery ticket themselves (treatment group). Subjects who had chosen their ticket themselves were, on average, demanding four times more money than those who were simply given their tickets when asked to name a price for which they would sell it. Those who chose the ticket themselves mistakenly assumed they would therefore be more likely to win, although the winners would be determined entirely by chance. This mechanism has been labeled the "illusion of control" [Langer, 1975], although critics of this concept have argued that it rather identifies a pattern of people overestimating their ability to predict future outcomes than of people overestimating their ability to control future outcomes [Presson and

Benassi, 1996]. Both interpretations are regarded as sufficient for the purpose of our current analysis.

4.2.4 The Role of Feedback in Overconfidence

Providing feedback is commonly regarded as an efficient learning mechanism in Markov Decision Processes in order to compute optimal behavior [van Otterlo and Wiering, 2012; Puterman, 1994; Bar-Shalom and Tse, 1974] and offers great value for deriving logical inference rules in accordance with Bayesian reasoning in machine learning [Kasneci et al., 2010; Bertsekas, 2019]. While these points may hold for cyber-physical systems without human decision-makers, humans are often observed to react to feedback differently. Contradicting Markovian assumptions, behavioral economic research constantly observes errors in updating the information set [e.g., Camerer, 1999; Charness and Levin, 2005]. This indicates that providing more feedback information to humans does not necessarily lead to better estimations and decisions. The importance of outcome feedback in counteracting certain behavioral biases, through continuous "monitoring of progress through a judgment-action-outcome loop" [Kleinmuntz, 1985, p.692] and offering corrective adjustments, has been pointed out before [Hogarth, 1981], but it may not apply equally to every type of bias.

While research studying the connection of high information supply to overconfidence is comparably scarce, in various instances a higher amount of feedback was not found to be able to improve behavioral rationality with regards to overconfidence, as it would be expected by theory. Both overconfidence and underconfidence were shown to persist in employees' choices of incentive schemes although receiving clear feedback revealing the optimal alternative for them [Larkin and Leider, 2012]. Experiments on sports outcome-estimation showed that overconfidence did not decrease but rather increase with additional information, as the subjects' confidence increases more than their accuracy does [Tsai et al., 2008]. Similarly, CEO overconfidence in forecasts was found to persist against corrective feedback [Chen et al., 2015] and venture capitalists were shown to be more overconfident when having access to more information [Zacharakis and Shepherd, 2001]. In general knowledge tasks that are related to each other, corrective outcome feedback has shown able to mitigate underconfidence but not overconfidence [Subbotin, 1996]. It turns out that feedback appears to affect behavioral biases rather ambiguously with overconfidence tending to largely persist, or even growing, in response to additional information.

4.2.5 Misperception of Random Sequences

With regard to sequential human decision processes, another potential threat of high information supply leading to biased decision-making is the danger of falling subject to the "hot hand fallacy". The study of this error in the assessment of past outcomes originates in observations from basketball. If a player repeatedly scored on his past throws, he is commonly judged to have a "hot hand" by his

teammates and the audience. Therefore, the player's likelihood of scoring on future throws is judged to increase by the audience, even though future throws are independent of past throws. In doing so, the player's objective probability of scoring is overestimated based on his past successes [Gilovich et al., 1985]. This hot hand fallacy results from misjudging statistically independent favorable events to be connected to one another, implying a positive autocorrelation between them. Repeated past positive outcomes are therefore erroneously expected to occur again in the near future, thereby overestimating their objective probability of occurrence. The hot hand fallacy is a vivid illustration of biased probability judgments through presumed representativeness of recent information. Misinterpretations of random sequences through such an extrapolation of recent outcomes into the future were empirically found to apply to several contexts. Regarding basketball, the findings of Gilovich et al. [1985] were supported by studies on betting market odds, although only small effects were found [Camerer, 1989; Brown and Sauer, 1993]. In a laboratory experiment, simulating a blackjack game, Chau and Phillips [1995] observed gamblers to bet more money after a series of wins than after a series of losses. In more sensitive contexts, individual decisions show signs of the hot hand fallacy as well. For instance, financial investors who had previous negative experiences with low stock returns, exhibit a decreased willingness to take financial risks in future investments [Malmendier and Nagel, 2011]. Attribution theory, especially the illusion of control mentioned above, may factor into the hot hand assumption, as subjects are more likely to attribute recent sequences with low alternation to human skilled performance, while sequences with high alternation are rather perceived as chance processes [Ayton and Fischer, 2004].

Before proceeding, we note that the hot hand fallacy's opposing bias, the so-called *gambler's fallacy*, presents the erroneous expectation of systematic reversals in stochastic processes. These are grounded in the belief that a small sample should be representative for the underlying population, i.e., that a Bernoulli random process should balance out even over just a few rounds [Rabin and Vayanos, 2009; Suetens et al., 2016].

4.3 Surveillance Drone Piloting Framework and Hypotheses

Our experimental design for analyzing behavioral biases in CPHSs builds upon a basic setup involving UAVs that interact with human operators, as discussed in Feng et al. [2016]. Partly autonomous UAVs are used for road network surveillance inside a network that connects multiple traffic junctions. The drone pilot's objective is to gain the maximum information about the traffic at all junctions using a single drone which follows a pre-designed path, graphically illustrated in Figure 13.

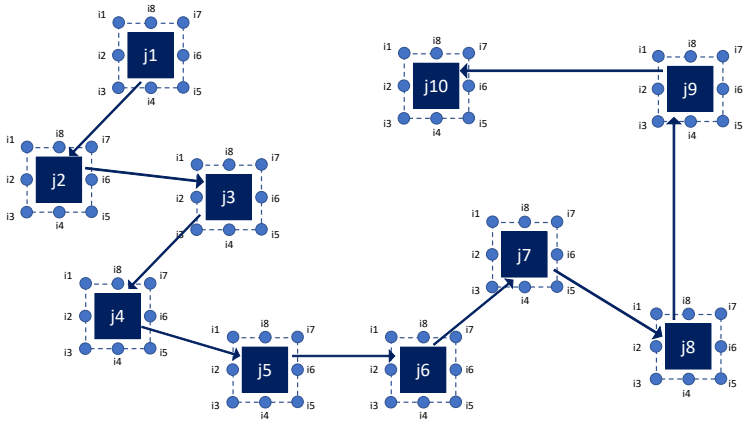


Figure 13: Illustration of a road network for UAV missions, adapted from Feng et al. [2016]. A single drone follows a given path overflying 10 junctions. At each junction $j \in \{1, 2, \dots 10\}$ a maximum of $N = 8$ rounds can be flown.

To increase the picture quality (and thereby the information obtained), the operator may choose to fly up to 8 rounds over each junction j, each round taking an additional photo. The individual photos are combined into a higher-resolution image, whose total value will be denoted I_j. This total value depends on the number of flights, say f_j, and on the random additional information content of each photo taken. The situation is modelled via the random process $\sigma_j(i)$ which quantifies the information content of the combined photo of junction j at round i. The process is initiated at $\sigma_j(1) = 25$ and through iteration of

$$\sigma_j(i+1) = \sigma_j(i) + s_j(i)\rho(\sigma_j(i)), \quad i = 1, 2, \dots, f_j - 1, \tag{1}$$

leads to $I_j = \sigma_j(f_j)$.

Here, the additional information content of the photo taken at round i is quantified by the concave function $\rho(\cdot)$ which describes a decreasing marginal yield:

$$\rho(25) = 25, \quad \rho(50) = 20, \quad \rho(70) = 10, \quad \rho(80) = \rho(85) = \rho(90) = \rho(95) = 5. \tag{2}$$

The above reflects the fact that each additional picture often comes with less information value added, since a rough image of the traffic flow is already gained by the earlier pictures, and subsequent pictures only help further sharpening the image. In (1), $s_j(i) \in \{0, 1\}$ is an i.i.d. Bernoulli random process, with probability $P[s_j(i) = 1] = p$. This models instances wherein the new picture offers no information value added over the previous one due to, for instance, bad weather causing poor visibility, strong wind or lacking flow of traffic.

Whilst taking more photos will often lead to combined images of higher quality and also with more

information content, after each individual flight there exists a (small) probability of r of the drone crashing, leading to a loss of D (the value of the UAV) and the inability to continue flying again over the current or later junctions.

Given the above, the value of the image obtained at each junction j belongs to the finite set $\{0, 25, 50, 70, 80, 85, 90, 95, 100\}$. The current combined value of the drone and images after flying i rounds at junction j satisfies

$$V_j(i) = Dc_j(i) + \sigma_j(i) + \sum_{\ell=1}^{j-1} I_\ell, \tag{3}$$

where

$$c_j(i) = \begin{cases} 1, & \text{if the drone is still intact after flying } i \text{ rounds at junction } j, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

The total value gained by the operator at the end of the mission is given by:

$$V = 400c_{10}(f_{10}) + \sum_{j=1}^{10} I_j. \tag{5}$$

4.3.1 Decision Problem

Given the possibility of the drone crashing, rather than simply intending to fly the maximum number of 8 rounds over each junction, the operator is faced with the decision of how many rounds to fly over each junction in order to maximize its profit V. This amounts to a sequential stopping rule problem defined over a finite horizon. To make decisions when to fly (or switch) to the next junction, in addition to knowing the system model and its probabilities, the operator receives feedback from the drone, see Figure14. Every time the UAV passes over a junction, it sends back $\sigma_j(i)$, the information gain obtained so far over the current junction.

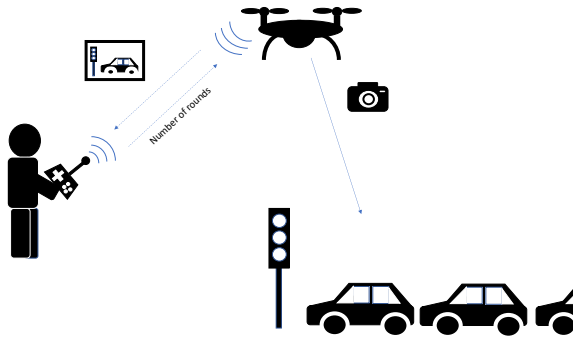


Figure 14: Human-in-the loop control system: The operator decides how many times the UAV should fly over each junction in order to gain accurate information about traffic conditions in the city. To assist in the decision-making, the drone may send feedback information about the picture quality, i.e., information gained.

Further the operator, is informed whether the drone has crashed. Using this information, the human operator is faced with the task of designing a closed loop flight policy.

The exact stopping rule can, in principle, be derived using dynamic programming [Bertsekas, 2005; Ferguson, 2008]. However, the value function depends not only on the value $\sigma_j(i)$, but also on the current round i as well as the junction j, with later junctions having less value. Thus, instead of pursuing an optimal strategy, a 1-stage look-ahead rule within the current junction becomes a reasonable alternative. Using such a myopic policy, the operator chooses to flight another round over the current junction j if and only if the drone is intact ($c_j(i) = 1$), $i \leq 8$ and the marginal gain of flying one more round is positive:

$$g(\sigma) \triangleq E[V_j(i+1) - V_j(i) \,|\, \sigma_j(i) = \sigma] = p\rho(\sigma) - Dr > 0, \tag{6}$$

where we have used (3), see also Figure15.

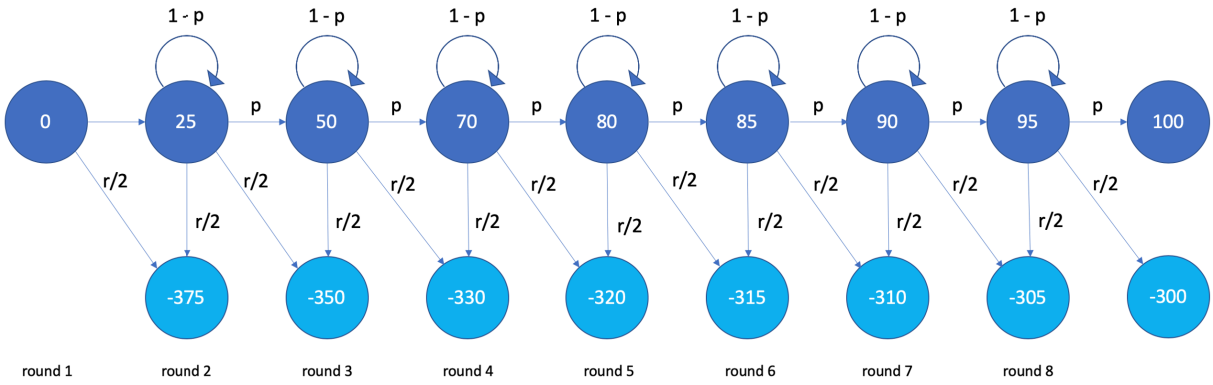


Figure 15: Marginal gains within each junction, for $D = 400$. Diagonal arrows describe an increase in information value with a subsequent crash, vertical arrows represent a non-increase with a subsequent crash. Note that the maximum information value gain attainable at each traffic junction, namely 100, can only be reached if a subject decides to fly all eight rounds possible while obtaining an increase in each round.

In the sequel we will fix the parameters to $D = 400$, $p = 0.5$ and $r = 0.02$, so that (2) provides

$$g(25) = 4.5,\ g(50) = 2,\ g(70) = -3,\ g(80) = g(85) = g(90) = g(95) = -5.5. \tag{7}$$

This makes flying until an information value $\sigma_j(i) = 70$ is accumulated in the current junction the optimal choice. From the fourth node onward, the marginal gain from flying another round is lower than the marginal loss, therefore a rational (and risk-neutral) agent should refrain from flying further rounds.

In addition to this full feedback case, we will also consider a situation wherein the drone does not transmit the feedback information $\sigma_j(i)$ to the operator. This requires the operator to design an open

loop policy [Bar-Shalom and Tse, 1974], as detailed below.

4.3.2 Experimental Drone Framework

We abstract the framework into an economic experiment (see Appendix C.1) in order to gain insights on actual individual behavior in human-in-the-loop control. The experiment translates the CPHS concepts of open versus closed loop control into a sequential decision-making game, framed in a context in which the subjects are owners to an UAV with a photo function. They are told that they were hired by the local city administration to support the city's traffic surveillance by taking photos of ten important traffic junctions. As outlined above, the subjects' task is to decide, how many rounds the UAV should fly over each of the 10 traffic junctions (up to a maximum of 8 times), while the drone would fly and take pictures autonomously. Subjects are incentivized, as they are paid according to the amount of information value gained through the pictures their drones take in the fictitious currency "Taler", with one Taler equaling one unit of information value gathered. Taler are exchanged into Euro, after all parts of the experiment are completed. However, during each round in which a UAV flies, it faces a constant crash risk of $r = 2\%$. At the end of this job, i.e., after all ten traffic junctions, subjects are able sell their drone for an additional fixed payment of $D = 400$ Taler, as long as it is still intact at this point, see (5).

Each traffic junction represents one stage of the experiment. For the sake of simplicity, subjects do not have to bear any costs from flying the drone. Further, the drone's value does not diminish from usage and subjects are told that batteries for the drones as well as (ground) transport between the traffic junctions is provided free of charge by the city administration. One battery charging allows the drone to fly up to eight rounds before the drone must land to recharge. Therefore, subjects are able to decide for a maximum of eight rounds per traffic junction and consequently take up to eight pictures. This corresponds to a maximum of eight nodes that can be reached per junction. In the first round flown over each traffic junction, an information value of $\sigma_j(1) = 25$ is gained with certainty, as any picture taken represents an increase in information value over no picture taken at all. From the second round onwards, the information value potentially gained per round decreases, see (2). For the sake of simplification, a crash can only happen after a picture is taken. As per (5), the information value gained at the junction where the crash occurs up to the point of it, still qualifies for the payment, as subjects were told that the drone's memory chip would survive the crash.

The operators need to decide how many rounds a UAV should fly over each traffic junction in order to maximize the total gain V in (5) and consequent payoff. In order to do so, some of them receive feedback on the results of the previous rounds – depending on the treatment – which they can base their decision on. As noted before, the resulting optimization problem belongs to the class of general

stopping problems. Its solution, both with and without feedback information, can in principle be derived, but requires careful computations. Since individuals are well-known to have limited cognitive and computational capabilities, they will not be able to compute a globally optimal strategy. Instead, optimizing subjects will use the decision heuristic in (6), i.e., the fourth node is identified as the one to reach.

The above leads to two decision heuristics, depending on whether feedback information is available or not. In the former (**closed loop heuristic**), at every junction the UAV should aim to fly as many rounds as needed until reaching an information value of $\sigma_j(i) = 70$. To be more precise, introduce $a_j(i) \in \{0, 1\}$, where $a_j(i) = 1$ corresponds to the decision to fly another round over the current junction, then the closed loop heuristic can be stated via:

$$a_j(i) = 1, \quad \text{if and only if } \sigma_j(i) < 70 \text{ and } i \leq 8 \text{ and the drone is intact.} \tag{8}$$

Note that in this case, the total number of rounds flown over a junction j, namely $f_j = \sum a_j(i)$, depends on the sequence of increases and non-increases the subject experiences.

In the absence of feedback from the drone, a suitable **open loop heuristic** amounts to attempting to fly $f_j = 5$ times at every junction:

$$a_j(i) = 1, \quad \text{if and only if } i \leq 4 \text{ and the drone is intact.} \tag{9}$$

This yields the expected result of three increases in information value, with the first one being granted with certainty in the first round. The above rules of thumb present reasonable heuristic approximations to the task's optimal solution when implemented for all junctions. A rational and risk-neutral decision-maker, who acts according to expected utility theory [von Neumann and Morgenstern, 1944], would follow these heuristics. Deviations from these decision-rules upwards or downwards represent indicators of overconfidence or underconfidence respectively, depending on the direction.

The experiment ends either after all junctions are completed and the drone is sold or once the drone has crashed. A crash could be caused by external factors like weather conditions, the drone hitting some object or animal, or technical failure. This risk of a total loss of the drone's value in case of a crash imposes the prospect of high one-time costs onto the participants and thus prevents them from simply choosing the maximum number of rounds possible at every traffic junction. As subjects are paid according to the information value gained, striving for the optimal information value (using the above heuristic) presents the rational strategy. Always flying the maximum number of 8 rounds is not efficient, as gaining information value has to be weighed against maintaining the chance of being able to sell the drone at the end of the experiment, as indicated in (5).

4.4 Hypotheses

The variety of behavioral phenomena and stylized facts regarding decision biases and errors in relation to overconfidence presented in the preceding section, suggests inefficient human decision-making. In fact, individuals are commonly observed to act overconfidently. In doing so, they tend to underestimate systematic risks, insufficiently process joint and conditional probabilities, self-servingly misattribute causal relationships and erroneously expect favorable outcome sequences to continue in the future. These aspects should be particularly relevant in our example, since the probability of the drone crashing may appear small for one specific round (2%), while the overall risk of the UAV crashing at any time over the course of the experiment is substantially higher. Modeling the crash probability as a Bernoulli process, the chance of a crash when attempting to fly the maximum number of rounds at each of the ten junctions would be $1 - 0.98^{80} > 0.8$. A narrow focus on the low single-decision crash probability [Read et al., 1999] could therefore lead subjects to underestimate the overall crash probability and consequently take higher risk. Providing feedback is commonly assumed to diminish these behavioral biases but known to not being able to entirely resolve them, and in some instances even aggravating them. Subjects may rather attribute positive feedback to their own decision performance to themselves instead of realizing that it is caused by chance. All these aspects point in the direction of expecting overconfident drone piloting. Manager overconfidence in economic decision-making can be summarized to originate in either overestimating expected cash flows or underestimating risk [Hirshleifer et al., 2012]. Translated to our design, the drone pilots' overconfidence could either result from overestimating the likelihood of information value improvements or from underestimating the risk of the UAV crashing. While the task is framed in a quite understandable context, it is non-trivial in its solution. Most subjects are expected to approach the task rather intuitively, i.e., not solving it systematically through comparing marginal gain and marginal costs, and consequently act overconfidently through one of its facets described above, if not through both. According to the general consensus of literature on overconfidence in human decision-making, the following hypothesis is posed:

Hypothesis 1: Drone pilots will generally act overconfidently in flying more than the optimal number of rounds, regardless of feedback.

As is also known from the literature, individuals are prone to unreflective repeat decisions that previously yielded them positive outcomes, even though such past successes entirely depended on chance and have no impact on future outcomes at all. This decision heuristic, the hot hand fallacy already discussed, presents a major misinterpretation of random sequences. While the hot hand fallacy in regard to the original study on basketball throws [Gilovich et al., 1985] is subject to criticism today

(with the data being re-analyzed and the conclusion being reversed [Miller and Sanjurjo, 2018b]), the general idea of individuals misjudging the meaning of random outcome sequences remains to be tested in the context of CPHS. For sequential decision problems featuring feedback from the system, decision-makers appear especially vulnerable to fall victim to the hot hand fallacy. Translated to our drone framework, it is therefore expected that those drone pilots who immediately reach the optimal information value of a junction through experiencing gains in repeated rounds in the closed loop system (as the possibility is only given there), perceive themselves to be "on a roll", i.e., as having an equivalent to a hot hand[20]. Since decision-makers are tempted to expect such apparent trends to continue, they will likely fly beyond the optimal number of rounds, making their operation inefficient with regards to risk and reward. This way, subjects fall victim to the hot hand fallacy[21] Compared to the gambler's fallacy, a hot hand is considered more suitable to portray drone controller's behavior in his attempt to gather information value, with humans more prone to expect favorable outcomes to repeat under their apparent control [Ayton and Fischer, 2004][22]. This leads to our second hypothesis:

Hypothesis 2: Sequences of immediate positive feedback will lead to drone pilots falling victim to the hot hand fallacy.

These two hypotheses were tested in an experiment involving students at Paderborn University in September 2019. Details are given next.

4.5 Experimental Design

In order to be able to compare behavioral effects in closed and open loop operation, a between-subject design experiment is conducted, meaning each subject can only participate in exactly one treatment with the groups being compared afterwards.

4.5.1 Treatments

To evaluate the effect of feedback in human-in-the-loop control, two treatments are presented: Closed Loop and Open Loop. The groups differ only in the fact that subjects in one of them receive immediate

[20] In our experiment, we define a hot hand as three consequent increases from the very start of a junction. With the junctions being distinct from each other and the first success being certain anyway, there is no possibility of pattern overlays [Miller and Sanjurjo, 2018a]. Therefore, as well as the fallacy being defined as the decision to attempt a fourth flight – regardless of its outcome – in our framework, we do not consider our results to be prone to a "streak selection bias" [Miller and Sanjurjo, 2018b].

[21] Literature observes a shift from hot hand beliefs for shorter outcome streaks towards a gambler's fallacy pattern for increasing streak length [Jarvik, 1951; Rabin and Vayanos, 2009]. Drone controllers who experience multiple increases in a row would therefore begin to overestimate the probability of a non-increase at some point and stop flying to prevent a crash accordingly.

[22] Also, a hot hand fallacy can be considered the more problematic error compared to the gambler's fallacy, since in the case of the latter, subjects would stop earlier and not risk a drone crash through flying in periods that yield marginal losses (although sacrificing marginal gain).

outcome feedback after every decision made, while subjects in the other group do not receive any feedback until the end of the experiment.

In the **Closed Loop Treatment**, subjects receive feedback directly after each round of each junction on whether an increase in information value was gained in this specific round, see Figure 15. They were also informed about the current total information value gained for the respective junction and whether the drone was still intact. Subsequently, they were asked whether they wanted to fly another round. The framing in the experimental instructions stated that the drone would transmit the pictures taken to the drone pilot's laptops immediately, so feedback was given just in time for the next round's decision.

In the **Open Loop Treatment**, subjects are provided with no feedback at all. Subjects make decisions on how many rounds the drone should fly over each junction entirely in advance. Afterwards, they are informed about the total information value they obtained and whether the drone is still intact and sold at the end of the experiment. The framing in the instructions states that the UAV has to be deconstructed in a complex procedure to be able to extract and read out the memory chip to check on the pictures taken. Therefore, no feedback on the pictures would be provided until either all junctions were completed or the drone had crashed.

4.5.2 Procedure

The experiment was computerized using the experiment software oTree [Chen et al., 2016] and hosted centrally on a university server, so subjects could remotely access the experiment and did not have to come to the laboratory physically. 500 subjects who previously had enrolled voluntarily into the BaER-Lab (`www.baer-lab.org`) student participant pool were chosen randomly, with 250 each being randomly assigned to the two treatments in advance. These subjects were contacted via the online recruitment system ORSEE [Greiner, 2015] and invited to participate within the following five days.

The invitation email included a hyperlink that directed the subjects to their respective treatment, where they received the detailed instructions for the experiment (see Appendix C.2). The instructions of both treatments only differed regarding the provision of feedback, as explained above. To be able to progress to the drone flying task, subjects had to correctly solve four control questions. The questions revolved around the central parameters of the experimental design in order to assure that the subjects had read and understood the instructions in its critical parts. Subsequently, subjects advanced to the drone flying task, in which they had to make their decisions through clicking single choice buttons (see Appendix C.3). Subjects would receive one unit of the fictitious currency Taler for each unit of information value gained as pay for flying the drone. During the whole experiment, one subject's payoffs did not depend on the decisions of any other subject. After completing the drone flying task,

a result screen presented the total information value accrued over all traffic junctions, the state of the drone, i.e., whether it was still intact and consequently sold to earn 400 additional Taler or not, as well as the corresponding payoffs the subject generated in Euros. Payoffs were translated from Taler to Euro at an exchange rate of 1 Euro per 120 Taler. Subjects were then asked to answer a standardized questionnaire, that included questions on demographics, perceived task difficulty and the subject's reasons for choosing the numbers of rounds the way they did. Afterwards, subjects were presented with the multiple price list by Dohmen et al. [2011] (see Table 44 in Appendix C.4), in order to measure their risk preferences.

Filling out the price list was incentivized, as for every fifteenth subject one out of the list's twenty rows was randomly selected and the payoffs that the subject's chosen alternative row yielded in this row was added to the his total payoffs. In case Alternative A had been selected by the subject, the fixed amount stated in the respective row was simply added to its payoffs. In case Alternative B had been selected, the lottery was automatically played out and either €0 or €30 were added to the subject's payoffs. After completing the price list, subjects were informed about whether the list came into effect for their payoffs, and in case it did which row had been selected. On a final screen, subjects were informed about their total payoffs in Euros and thanked for their participation.

A subject's total payoff function can consequently be formalized as

$$\Delta m = \frac{1}{120} V + d(\ell)_{i=15x} \tag{10}$$

where V is the total value, as in (5). The term $d(\ell)$ represents the additional payoff from the multiple price list, conditional on the selected row ℓ of the list that a subject was paid in case of being a fifteenth participant, i.e., a subjects participant's ID being a multiple of 15.

Subjects were able to collect their payoffs in cash at the Chair's secretariat following their participation by stating a unique eight-figure ID-Code they had to create in the beginning of the experiment. That way the correct payoffs could be handed out to each subject, while maintaining anonymity about the subject's decisions made.

4.6 Experimental Results

Out of the 105 subjects who finished the online experiment, 57 participated in the Closed Loop and 48 in the Open Loop treatment respectively. The average age of the subjects was 23.9 years and varied by about half a year between treatments. 31 subjects (29.52%) were female, while one subject classified itself as non-binary. On average, subjects earned €4.89 from the experiment, plus the amount they would earn from filling out the multiple price list in case they were paid because they were a fifteenth

participant.

Subjects in the Closed Loop treatment flew, on average, 5.89 rounds over each traffic junction, thereby exceeding the average number of rounds from the Open Loop treatment by 0.79 rounds per junction. Over all rounds flown, subjects in the Closed Loop treatment gained a cumulative information value of 651.84, on average. This represents a 27.83% surplus compared to the average information value gained in the Open Loop treatment. Similarly, a higher drone crash rate of nearly 70% was observed in the Open Loop group. These two dimensions generally interact, since the information value is fixed once a drone crashes. Averages of rounds flown and accumulated information value as well as average earnings and crash rates for each treatment are displayed in Table 14.

Table 14: Average Values of Rounds flown, gained Information Value V, Earnings Δm and Crash rates, by Treatment.

	Closed Loop	**Open Loop**
Average rounds flown	5.89	5.10
	(2.08)	(1.82)
Average total information value gained	651.84	509.90
	(397.56)	(380.13)
Average earnings (in €)	5.43	4.30
	(3.31)	(3.17)
Crash rate	52.63%	68.75%

Note: Standard deviations presented in parenthesis.

As could be expected from control theoretic results [Bar-Shalom and Tse, 1974], the Closed Loop system is clearly observed to be more effective than the Open Loop piloting system, with regards to accumulating information about traffic conditions in our framework. Considerably more information value was aggregated with a closed loop of immediate outcome feedback being implemented, which also translates to a higher monetary payoff for the subjects. However, subjects displayed behavioral biases as we will see. Indeed, the average number of rounds in the Closed Loop treatment being nearly six already hints that several subjects tended to fly beyond the heuristically optimal number of rounds. In doing so, they would have taken inefficiently high risks, since marginal losses exceeded marginal gains in later rounds of each junction. For instance, instead of aiming to add five or ten more units of information value when already sitting at 70, subjects should stop and continue with the next traffic junction to obtain twenty-five units of information value with certainty at the same crash risk.

4.6.1 Overconfidence and Underconfidence

To determine whether a subject acts overconfidently, optimally or even underconfidently at a certain traffic junction, the number of rounds flown is compared to the heuristics discussed above, see (6). If a subject flew beyond the heuristically optimal number of rounds, the decision at this junction is

noted as overestimation. If the subject decides to fly less rounds than optimal the junction's decision indicates underestimation. In case the heuristic is met, the subject acts optimizing.

As mentioned before, flying until the fourth node is reached (associated with an information value of 70) is a suitable heuristic for a risk neutral decision-maker. This induces different strategies for the treatments. In the closed loop system, it is optimal for the subjects to fly as many rounds as necessary to reach the fourth node. The actual number of rounds needed depends on the sequence of increases and non-increases experienced, as illustrated in Figure 15. In the open loop system, meanwhile, flying five rounds is a suitable heuristic. As subjects have to decide on the number of rounds to fly for all traffic junctions in advance, this strategy applies to all of them, as explained before.

An indicator for overconfidence on the subject-level is created by calculating the quotient of the number of overconfident junctions and the total number of junctions per subject, since the number of traffic junctions played differed between subjects due to some drones crashing prematurely. This relative frequency of overconfident junctions by a subject will be labeled "overconfidence degree". Degrees of underconfident and optimal decisions are computed analogously, so all three degrees add up to 1. The average degrees of each of these behavioral tendencies are displayed in Table 15.

Table 15: Average shares of overconfident, underconfident and optimal decisions

	Closed Loop	**Open Loop**
Overconfident	0.441	0.417
	(0.36)	(0.35)
Underconfident	0.309	0.383
	(0.31)	(0.33)
Optimal	0.252	0.200
	(0.26)	(0.24)

Note: Standard deviations presented in parenthesis.

On average, in the closed loop system, subjects make overconfident decisions for 44% of the traffic junctions. Comparatively, shares of overconfident decisions do barely differ between the two treatments with subjects deciding overconfidently for approximately 42% of the junctions in the open loop system. The average share of underconfident decisions was higher in the Open Loop treatment. Shares of overconfident and underconfident decisions are relatively close to each other for the open loop, while overconfidence characterizes the predominating behavioral tendency in the closed loop system. The share of optimal decisions drops off in both treatments, with subjects only deciding optimally at one fourth and one fifth of the junctions, respectively. This gap also becomes apparent in the graphic illustration of decision patterns in Figure 16. Meanwhile, the shares of overconfident, underconfident, and optimal decisions did not differ significantly by gender in either treatment.

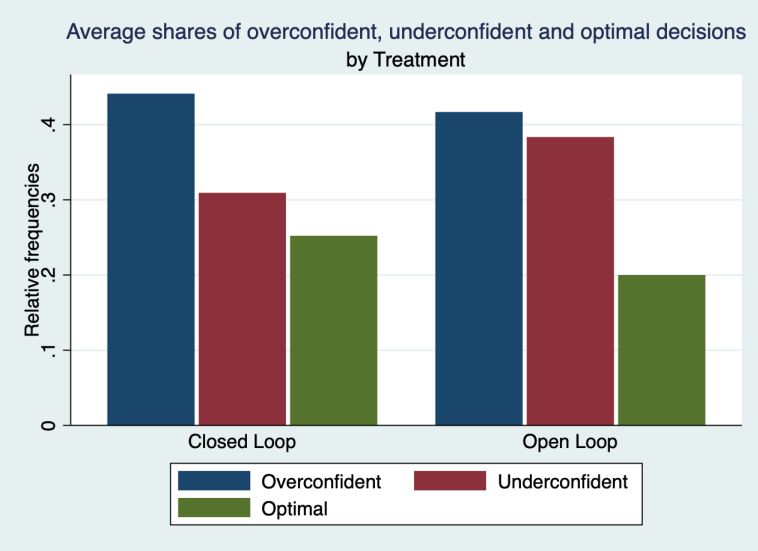


Figure 16: Average shares of overconfident, underconfident and optimal decisions

Seven subjects in the Closed Loop and eight subjects in the Open Loop group behaved overconfidently in every single phase they flew in, while only three subjects in the closed loop system and two in the open loop system decided optimally for every junction, while three subjects in each treatment acted underconfidently at every junction. Among those junctions in which subjects act overconfidently, they flew, on average, 2.88 (SD = 1.36) more rounds after reaching the optimal node in the Closed Loop treatment. In the Open Loop group, subjects flew on average 1.88 (SD = 0.85) rounds beyond the optimum, respectively. It appears that those subjects, who decide to fly more rounds than optimal, do it clearly, i.e., usually by more than one round. Also, subjects exceed the optimal number of rounds by, on average, one more round in the closed loop compared to the open loop system once they fly beyond the optimum. The difference is statistically significant in a Mann-Whitney U-Test ($|z| = 7.393, p = 0.000$) and matches the observation of subjects in the Closed Loop treatment flying more rounds overall and accruing more information value in total.

To gain a more precise impression about the distribution of subjects' deviations from optimal behavior within each treatment, we classify all subjects into broad categories. Subjects that act in accordance with the strategy of flying until reaching the fourth node at all traffic junctions are classified as *optimizing*. A subject whose overconfidence degree exceeds one third, up to 0.5, is classified as *rather overconfident*. Once a subject flies beyond the optimal number of rounds in more than half of the junctions (i.e., displaying an overconfidence degree over 0.5), this subject is considered *strongly overconfident*. Analogously, the categories *rather underconfident* and *strongly underconfident* are defined using the same thresholds in terms of underconfidence degrees as those for overconfidence. The sixth category, *mixed*, defines scenarios, in which subjects show overconfident or underconfident behavior in at least one junction but in less than one third of all junctions they played (equaling

overconfidence or underconfidence degrees between 0 and 0.333). Since these subjects act optimizing in some junctions but not in all, their behavior still contradicts the notion of a strictly optimizing agent, although no clear dominant decision-pattern can be assigned to this behavior. The distribution of subjects into these categories by treatment is displayed in Table 16.[23]

Table 16: Categorization of Decision Behavior, by Treatment

	Closed Loop	**Open Loop**	Total
Optimizing	3	2	5
	(5.36%)	(4.26%)	(4.85%)
Rather overconfident	5	6	10
	(8.93%)	(12.77%)	(10.68%)
Strongly overconfident	24	16	40
	(42.86%)	(34.04%)	(38.83%)
Rather underconfident	8	5	13
	(14.29%)	(10.64%)	(12.62%)
Strongly underconfident	13	16	29
	(23.21%)	(34.04%)	(28.16%)
Mixed	3	2	5
	(5.36%)	(4.26%)	(4.85%)
Total	56	47	103
	(100%)	(100%)	(100%)

Note: Relative shares referring to each treatment are presented in parenthesis.

As shown in Table 16, only few subjects in each treatment act optimally throughout. Rather, individual behavior deviates from the optimal number of rounds in both directions. Approximately half of the subjects in both groups (Closed Loop: 51.79%, Open Loop: 46.81%) can be considered rather overconfident or strongly overconfident. Additionally, large shares of rather underconfident or strongly underconfident individuals (Closed Loop: 37.50%, Open Loop: 44.68%) are observed as well. From the assumptions of MDPs and standard economic theory follow notions of individuals as rationally optimizing decision-makers that leave no room for overconfident or underconfident behavior. Consequently, our empirical results do not meet these theoretical propositions. In fact, according to a Binomial Test of the data for the Closed Loop treatment the probability of observing 29 overconfident (rather or strongly overconfident) subjects out of a sample of 56 is virtually zero ($p < 0.000001$), given the assumption of optimizing behavior. The same result is obtained for overconfidence in the open loop system. Consequently, we can fully support **Hypothesis 1**, indicating a strong general tendency of individuals to act overconfidently in CPHSs, regardless of whether receiving immediate outcome feedback or not. While not part of the hypothesis, we also observe the equivalent results for underconfident behavior in both treatments.

[23]Two subjects from the Closed Loop treatment were excluded, because they crashed in the very first round at the first traffic junction, wherefore no statement on their decision-making behavior can be made.

4.6.2 Hot Hand Fallacy

In contrast to an open loop, only a closed loop system provides the threat of sequences of immediate outcome feedback luring subjects into a hot hand fallacy. In our experiment, a subject in the Closed Loop treatment is defined to have a "hot hand" once it reaches the optimal node through obtaining three increases in information value in a row even though the first increase was certain. This corresponds to the common perception of three repeated outcomes as a streak [Carlson and Shu, 2007]. In this case, an optimizing subject would stop flying when following the one-stage optimal strategy, since three increases are necessary to reach the optimum and flying beyond this optimum yields a negative marginal gain. The outcome of prior rounds should have no relevance for the subject's decision, since the rounds' outcomes are independent of each other. If a subject decides to continue flying though, it violates the heuristically optimal decision rule, falling victim to the hot hand fallacy.

Overall, at 84 junctions the respective subject experienced three continuous increases in information[24]. In 70 out of these 84 cases (83.33%), subjects decided to fly at least one more round, thus falling victim to the hot hand fallacy. Under the assumption of subjects being optimizing Markov decision-makers, the probability of such a high proportion disobeying the optimal strategy is practically zero (Binomial Test: Pr(all subjects act optimally | 83.33% of the hot hand situations result in the hot hand fallacy): $p < 0.0001$). Consequently, **Hypothesis 2** is strongly supported.

Comparing non-overconfident and overconfident decision-making between junctions with and without a hot hand in Table 17, we observe those subjects who experience a hot hand situation (70 out of 84, 83.33%) to have a significantly higher chance of acting overconfidently at that junction, compared to those subjects who do not reach the optimum through three consecutive gains in information value (123 out of 260, 47.31%).

Table 17: Hot Hand decision-making

	Not overconfident	Overconfident	Total
No hot hand	137	123	260
Hot hand	14	70	84
Total	151	193	344

Note: The table compares outcome efficiency between instance resembling a hot hand and others for any traffic junction in the Closed Loop treatment.

Running a Chi-Squared Test to test the hypothesis statistically, we obtain a highly significant relationship between hot hand situations and subsequent overconfident behavior, i.e., the hot hand fallacy, with an error probability p of virtually zero ($\chi^2(df = 1) : 33.46$, $p < 0.0001$).

[24]The analysis for this Section is conducted at the decision level rather than the subject level.

4.6.3 Risk Preferences

Subjects' individual risk preferences are broadly classified to be risk neutral, risk averse or risk seeking (for a more detailed explanation see Appendix C.4). Given that the presented strategy is based on the assumption of a risk neutral decision-maker, the observed distribution of risk preferences in each treatment is displayed in Table 18[25].

Table 18: Distribution of Risk Preferences, by Treatment

	Closed Loop	**Open Loop**	Total
Risk averse	36	33	69
	(63.16%)	(68.75%)	(65.71%)
Risk neutral	6	5	11
	(10.53%)	(10.42%)	(10.48%)
Risk seeking	6	4	10
	(10.53%)	(8.33%)	(9.52%)
Not identifiable	9	6	15
	(15.79%)	(12.50%)	(14.29%)
Total	57	48	105
	(100%)	(100%)	(100%)

Note: Percentage values in parenthesis display relative frequencies of risk preferences in the respective treatment.

Consistent with the literature, the clear majority of around two thirds of the subjects in both treatments displays risk aversion. Overall, the distribution of risk preferences appears very similar between the treatments, with the relative shares of each risk attitude varying only slightly between the groups. This impression is supported by a two-sided Kolmogorov-Smirnov test, in which the null hypothesis that the risk preference distribution in the two treatments is not statistically different from each other cannot be rejected ($D = 0.030$, $p > 0.1$). Also, degrees of overconfidence and underconfidence did not differ significantly in between-treatment comparisons for each risk attitude using Mann-Whitney U-tests.

Further, we did not observe risk seeking subjects to display significantly more overconfident behavior (Open Loop: $\varnothing = 0.5257$, Closed Loop: $\varnothing = 0.175$) than risk neutral (Open Loop: $\varnothing = 0.4333$, Closed Loop: $\varnothing = 0.700$) or risk averse (Open Loop: $\varnothing = 0.4140$, Closed Loop: $\varnothing = 0.3818$) individuals in either treatment (Kruskal-Wallis equality of populations tests, Closed: $H = 0.450, p = 0.7984$, Open: $H = 5.437, p = 0.0660$)[26]. On the other hand, in the closed loop system, we find risk averse subjects to be significantly more underconfident, compared to the risk neutral and risk seeking subjects ($\varnothing$ underconfidence degree: 0.3480 $SD = 0.3018$), Kruskal-Wallis equality of populations test: $H = 8.891, p = 0.0117$), while no effect was found in the open loop system.

[25] Risk preferences of fifteen subjects in total could not be determined, since their choices switched between Option A and Option B more than once.

[26] Note that statistical tests on risk preferences have to be interpreted with caution, since the sample size of risk neutral and risk seeking individuals is very low.

Besides overconfidence, risk seeking preferences could in theory present an alternative explanation for drone pilots flying beyond the optimal number of rounds, with subjects primarily aiming to achieve a large information value while just hoping for their drones not to crash. However, we do not find substantial differences in overconfidence conditional on risk attitude between treatments and overall mainly observe individuals who can be classified as risk averse. This would theoretically predict subjects to decide rather conservatively through settling for a lower number of rounds in order to remain in the game and to protect the drone from crashing – even at the cost of missing out higher payments from information value gains. Since the experimental results show a large share of subjects flying beyond the optimum, the case for individual overconfidence as the dominant explanation for inefficient drone piloting beyond the optimum is strengthened. Consequently, we conclude that risk preferences are not able to explain overconfident behavior, while they might explain the observed degrees of underconfidence).

4.7 Discussion and Conclusion

Our experimental drone framework allows to observe how individuals behave when faced with the task of piloting an UAV under risk and uncertainty, paralleling a real-world decision-problem. Even though a closed feedback loop was identified to be the more successful system compared to open loop operation, we still observed inefficient drone piloting from the vast majority of subjects. Individuals expressed both overconfident and underconfident behavioral tendencies, regardless of receiving immediate outcome feedback. Specifically, our results indicate that immediate outcome feedback, that is originally intended to support optimal decision-making, can turn out to be rather counterproductive in this regard. Overconfident decisions and consequently inefficient drone piloting can be facilitated by the hot hand fallacy as a misinterpretation of random sequences of immediate feedback on positive outcomes, since subjects fail to realize such sequences to be caused by chance and therefore being history independent. In fact, a handful of subjects in the closed loop treatment stated in the questionnaire that their decision strategy was to fly as long as they achieved steady increases in information value, trying to exploit an apparent hot hand. The fact that the possibility for this fallacy is only given in a closed loop system, presents an obvious weakness that should be considered in designing such feedback policies.

In general, the current work exposes the human as an under-observed source of errors in human-in-the-loop control systems. We thus advise researchers and practitioners to carefully account for the behavioral component in the control of cyber-physical systems and the potential problems that arise from it, besides mathematical model optimization only. In particular, our findings illustrate the impact of behavioral biases regarding effects of immediate feedback and the (miss-)understanding of history

independence in chance processes.

While more information is commonly regarded to result in better decisions in cyber-physical systems, human susceptibility for perceptual biases in response to high information supply has to be taken into account. Therefore, identifying an optimal quantity and frequency of feedback remains a goal for future research. We expect a carefully crafted intermittent feedback to be better suited for this purpose and stress the need for an intelligent feedback design that adapts to an individual's rationality in order provide suitable amount of information.

Generally, considering the effect of humans in control loops more seriously presents an important issue for research and practice. Overall, humans were shown to mostly not act optimizing in the given decision-problem in our experimental framework, which strongly puts the Markov decision-maker as an adequate characterization of human decision-making in question. Models of human decision processes should be revisited to account for limits of cognitive capacities and behavioral biases that result from them, in order to not jeopardize technological accomplishments through erroneous human decisions. Otherwise, individuals in human-in-the-loop control might take unnecessarily high risk and render thoughtfully designed policies inefficient, as seen for highly frequent feedback in the case of the hot hand fallacy.

Lastly, our study further provides a methodological contribution to research on CPHSs, making a first approach to incorporate insights from behavioral economics into control engineering. Further it introduces incentivized economic experiments as a viable option to reveal how individuals actually behave, in contrast to how they are theoretically prescribed to behave. We present an experimental UAV framework featuring a sequential decision-problem, with a focus on behavioral biases in relation to feedback policies. Future experimental research in this area may intensify efforts to incorporate various other behavioral phenomena and stylized facts into control engineering by building upon this framework, in order to design or test behavioral interventions that are able to proactively counteract them.

Chapter 5:

"Explaining Apparently Inaccurate Self-Assessments of Relative Performance: A Replication and Adaptation of 'Overconfident: Do you put your money on it?' by Hoelzl & Rustichini (2005)"

Explaining Apparently Inaccurate Self-Assessments of Relative Performance: A Replication and Adaptation of "Overconfident: Do you put your money on it?" by Hoelzl & Rustichini (2005)

Marius Protte*,†

Abstract

This study replicates and adapts the experimental framework of Hoelzl and Rustichini (2005), which examined overplacement – overconfidence in relative self-assessments – by analyzing individuals' voting preferences between a performance-based and a lottery-based bonus payment mechanism. The original study found underplacement – the majority of their sample apparently expected to perform worse than others – in difficult tasks with monetary incentives, contradicting the widely held assumption of a general human tendency toward overconfidence. This paper challenges the comparability of the two payment schemes, arguing that differences in outcome structures and non-monetary motives may have influenced participants' choices beyond misconfidence. In an online replication of their experiment, a fixed-outcome distribution lottery mechanism with interdependent success probabilities and no variance in the number of winners – designed to better align with the performance-based payment scheme – is compared against the probabilistic-outcome distribution lottery used in the original study, which features an independent success probability and a variable number of winners. The results align more closely with traditional overplacement patterns than underplacement, as nearly three-fourths of participants prefer the performance-based option regardless of lottery design. Key predictors of voting behavior include expected performance, group performance estimations, and sample question out- comes, while factors such as social comparison tendencies and risk attitudes play no significant role. Self-reported voting rationales highlight the influence of normative beliefs, control preferences, and feedback signals beyond confidence. These results contribute to methodological discussions in overconfidence research by reassessing choice-based overconfidence measures and exploring alternative explanations for observed misplacement effects.

JEL Classification: D91, D82, D83, C90, C18

Keywords: Overconfidence, overplacement, performance, self-assessment, lottery design

*Paderborn University, Heinz-Nixdorf-Institute, Fürstenallee 11, 33102 Paderborn

†Corresponding author, marius.protte@upb.de

This article is currently under review at *Journal of Economic Psychology*.

Preprint available at arXiv: https://doi.org/10.48550/arXiv.2507.15568

5.1 Introduction

Overconfidence was famously labeled the "[perhaps] most robust finding in the psychology of judgment" by DeBondt and Thaler [1995, p. 389], and is frequently cited as a key driver of inefficiencies in human decision-making and resulting economic outcomes. The related literature attributes a wide range of financial, managerial, and societal phenomena – at least in part – to humans' overconfidence regarding their skills, capabilities, judgments or prospects. In financial markets, overconfidence has been linked to excessive trading volume and heightened market volatility [Odean, 1998; Daniel et al., 2001; Statman et al., 2006], with diminished trading performance [Biais et al., 2005], overreactions to private information and underreactions to public information, increasingly aggressive trading after past gains, underestimation of risk, and investment in riskier securities [Chuang and Lee, 2006]. In corporate finance and management, CEOs identified as overconfident are more likely overinvest in internal funds [Malmendier and Tate, 2005], pursue mergers and acquisitions at an elevated rate [Malmendier and Tate, 2008], and engage in self-serving attributions of company performance [Libby and Rennekamp, 2012]. Overconfidence has also been linked to the emergence of speculative price bubbles [Scheinkman and Xiong, 2003], premature or excessive market entry in highly competitive industries [Camerer and Lovallo, 1999; Cain et al., 2015; Vörös, 2024], and a reluctance to make necessary strategic adjustments [Kraft et al., 2022; Gervais and Odean, 2001], ultimately increasing the risk of entrepreneurial failure [Bernardo and Welch, 2001; Simon et al., 2000][27]. Beyond financial and managerial contexts, overconfidence has been implicated in various labor market outcomes [Santos-Pinto and de la Rosa, 2020] and broader social and political dynamics. It has been associated with self-deception in decision-making [Bénabou and Tirole, 2002], under-insurance and inadequate risk protection [Sandroni and Squintani, 2007], failed marriages [Mahar, 2003], and the outbreak of wars [Johnson and Tierney, 2011].

While early research – explicitly or implicitly – treated overconfidence as one uniform psychological concept [Moore and Healy, 2008], a more nuanced understanding has since been established. Typically, literature cites three specific phenomena that are amalgamated under the term "overconfidence": overestimation, overplacement, and overprecision. Regarding one's actual ability, performance, control, or chance of success, *Overestimation* describes a self-assessment that exceeds the level that is objectively justified; *Overplacement* means exaggerating one's placing relative to a reference population, implying a belief to be better than others; *Overprecision* refers to excessive certainty regarding the accuracy

[27] At the same time, (over)confidence is considered a major driver of success in investing [Wang, 2001], attaining positions of influence [Anderson et al., 2012; Radzevick and Moore, 2011], becoming an entrepreneur [Townsend et al., 2010] and fostering innovation [Baek and Neymotin, 2019]. These associations may be explained by overconfident individuals appearing more competent than those who are objectively competent [Anderson et al., 2012], them having propensity to exert greater effort [Landier and Thesmar, 2003], or greater willingness to take the necessary risks to realize desired outcome in competitive contexts with overconfidence counteraction inefficient risk-aversion [Kahneman and Lovallo, 1993], enabling individuals to pursue opportunities they might otherwise forgo.

of one's beliefs [Moore and Healy, 2008][28]. The type of individuals' inaccuracy in self-assessments is typically observed to be confounded with the difficulty of the task on which the assessment is based. Multiple studies show that overplacement occurs on easy tasks, while underplacement occurs on difficult tasks, [Healy and Moore, 2007; Moore and Cain, 2007; Moore and Small, 2007; Moore and Healy, 2008; Grieco and Hogarth, 2009]. The interrelation between overestimation and underestimation runs counter-directional, with overestimation occurring on difficult tasks and underestimation on easy ones [Healy and Moore, 2007; Grieco and Hogarth, 2009]. Also, low performance in absolute terms is commonly found to facilitate underplacment on difficult tasks, while high performance are found to be associated with overplacement [Kruger, 1999; Brown et al., 2016]. Meanwhile, estimation and placement are commonly observed to be negatively correlated with respect to task difficulty. Overplacement typically occurs on easy tasks, while overestimation takes place on difficult tasks [Healy and Moore, 2007; Moore and Healy, 2008; Grieco and Hogarth, 2009]. Overprecision is positively correlated to both other overconfidence types in both directions, i.e. the more precise a self-assessment the less over- or underestimation and over- or underplacement occurs [Moore and Healy, 2008]. Overplacement[29] – and underplacement by association – are the paradigms of interest for the scope of this study.

A central effort for distinguishing under which circumstances overplacement and underplacement occur respectively was made by Erik Hoelzl and Aldo Rustichini in 2005. They introduce a novel experimental approach to studying overplacement under economic incentives and varying task difficulty in their well-cited original research paper "Overconfident: Do you put your money on it?". By letting participants vote between a performance- based (the best performing half in a knowledge test wins) and a lottery-based (half of potential outcomes in an individual die roll wins) bonus payment mechanism – both offering an 50% win probability in expectation — they use this choice as a behavioral measure for inaccurate self-assessments, with significant miscalibration in either direction being interpreted as overconfidence or underconfidence respectively. This design goes beyond verbal self-assessments of individuals' own skills and abilities, which were widely used in prior experiments despite being prone to subjective interpretation [e.g., driving skills in Svenson, 1981]. Notably, H&R observe a tendency toward underplacement rather than overplacement when a task is difficult and real (as opposed to hypothetical) bonus payments are provided. This finding contrasts with the prevailing assumption of a general and robust human tendency toward overconfidence, as expressed by DeBondt and Thaler [1995], among others. Since H&R, numerous experimental studies on overconfidence have adopted

[28]Moore and Healy [2008] criticize researchers' habits of treating overestimation, overplacement, and overprecision as "interchangeable manifestations of self-enhancement" (p. 503) despite them being "conceptually and empirically distinct" (p. 515) types of overconfidence.

[29]Overplacement [Larrick et al., 2007] is closely related to the "better-than-average effect" [Alicke and Govorun, 2005]. The two terms are often used interchangeably in the literature [Moore and Healy, 2008] and will be treated as such in this study, for the sake of conciseness. However, hereafter only "overplacement" will be used, as it is the broader term for inaccuracies in relative self-assessments.

and refined similar performance-based versus conditional or random payment approaches to studying overplacement [see e.g., Blavatskyy, 2009; Urbig et al., 2009; Grieco and Hogarth, 2009; Park and Santos-Pinto, 2010; Ericson, 2011; Benoît et al., 2014; Owens et al., 2014; Koellinger and Treffers, 2015; Hollard et al., 2016; Benoît et al., 2022].

Meanwhile, legitimate skepticism has been raised about whether misconfidence can actually be inferred from over- or underplacement data, or from choice behavior in general [Benoît and Dubra, 2011]. Benoît et al. [2014] highlight the mixed evidence on overplacement and emphasize its already limited capacity to support generalizable conclusions within the traditional paradigm, irrespective of the criticism by Benoît and Dubra [2011]. The debate revolves around three main aspects. First, overplacement and underplacement observed in ranking experiments that rely on a single reference point – typically the median – and employ a dichotomous categorization of success to infer miscalibration in relative abilities can often be rationalized [Benoît and Dubra, 2011]. This applies to H&R's findings as well. Consequently, it is argued that such experiments reveal only "apparent", rather than "true", overconfidence or underconfidence. This distinction is particularly important, as true overconfidence may lead to the aforementioned negative real-world consequences, whereas apparent overconfidence is unlikely to have such effects [Benoît and Dubra, 2011; Benoît et al., 2014]. Second, overconfidence may be better understood as a statistical bias stemming from information asymmetries or limited information availability rather than as a behavioral or psychological bias [Healy and Moore, 2007; Grieco and Hogarth, 2009; Benoît and Dubra, 2011]. Third, alternative decision motives – such as a preference for control – may cause overplacement data in choice experiments where individuals bet on their own performance versus random devices, which is misinterpreted as overconfidence [Owens et al., 2014; Benoît et al., 2022]. Nevertheless, some studies [e.g., Merkle and Weber, 2011; Benoît et al., 2014] observe behavioral patterns that appear overconfident despite fulfilling Benoît and Dubra [2011]'s criteria for identifying 'true' overconfidence. Thus, prior overconfidence research should not be automatically dismissed or deemed invalid [Merkle and Weber, 2011]. As this debate remains ongoing and extends beyond the scope of this study, key aspects of the discussion will be further outlined in Appendix D.1. This study primarily aims to evaluate the internal validity of H&R's experimental results concerning the underlying motives behind subjects' voting behavior and self-assessment accuracy, rather than to assess their interpretation within the broader overconfidence paradigm. However, this replication will apply due caution in both the use of terminology and the interpretation of results, acknowledging the limitations outlined by the aforementioned research.

H&R infer underconfidence from a majority of participants preferring the lottery in their *Difficult x Money* treatment condition. Their identification strategy is based on the assumptions of individuals having an incentive to truthfully reveal their self-assessment relative to the group by voting the option

they believe maximizes their chance of receiving a bonus: selecting the performance-based payment scheme if they expect to rank in the upper half and opting for the luck-based die roll if they anticipate performing below the group median. Despite H&R's surprising results and their frequent citations, little attention has been paid to the characteristics of the lottery itself, even though the issue of alternative explanations for voting behavior has been discussed in the literature. This study argues that the two payment scheme alternatives are not fully comparable, as they differ in their degree of outcome dependency and allow for inconsistent numbers of participants receiving a bonus. Under monetary incentives and difficult tasks, these structural differences may alter voting behavior through mechanisms such as aversion to social comparison, low self-efficacy, prosociality or other non-monetary motives, potentially shifting votes in favor the lottery option. Therefore, this replication examines the robustness of H&R's results to a modified lottery mechanism that better aligns with the characteristics of the performance tests. Additional controls explore whether miscalibrated voting distributions, in fact, stem from inaccurate self-assessments or other underlying factors.

The importance of replications in experimental economic research – much like in social sciences in general – has been emphasized since the field's inception [see e.g., Smith., 1970; Rosenthal, 1990; Lamal, 1990; Charness, 2010]. By replicating H&R's experiment and examining whether the characteristics of the lottery device themselves may influence behavior, this study contributes to the ongoing methodological discussions on the feasibility of using performance-based versus random-device-based choices in experimental overconfidence research. It also contributes to developing a more comprehensive understanding of misplacement in individual self-assessments and behavioral factors besides misconfidence that may be mistakenly interpreted as such. Beyond methodological implications, the findings may also have practical relevance for employees' compensation plan choices and organizations' compensation plan design [Brown et al., 2016].

The replication results show voting distributions that align more closely with the traditional understanding of overplacement than with the underplacement observed in H&R. The alternative lottery mechanism does not appear to influence voting behavior. Additionally, four key factors driving individuals' voting decisions are identified: confidence level, signals obtained through sample questions, normative beliefs, and – exclusively among test voters – a preference for control.

The remainder of this paper is structured as follows: Section 5.2 provides a brief summary of the original study by H&R, outlining its research question, contribution, experimental design, and key findings. Section 5.3 then discusses the limitations of H&R's study and derives the hypotheses for the present replication study. Subsequently, the experimental design and experimental results will be presented in detail, following the structure of H&R's paper with selective modifications. Section 5.4 introduces task selection, sampling, experimental procedure, treatment variations, and the incentive

scheme. Section 5.5 presents descriptive and non-parametric analyses of the experimental results, examining subjects' predictions, voting behavior, and identifying predictors and correlates of their voting decisions. Additionally, an exploratory examination of subjects' self-reported voting rationales will be conducted. Finally, Section 5.6 summarizes the study's key findings and concludes with an in-depth discussion of its limitations and implications.

5.2 Original Experiment by Hoelzl & Rustichini (2005)

Instead of relying on verbal self-reports to measure subjective beliefs, as was commonly practiced at the time, Eric Hoelzl and Aldo Rustichini (hereinafter "H&R") present a novel approach to studying overconfidence in relative comparisons, i.e. overplacement. Their original experiment introduces a voting decision with payoff implications, serving as a more objective behavioral measure of overplacement relative to a reference group. The experimental design required subjects to indicate their preference between two distinct bonus payment allocation mechanisms – a performance-based one and a lottery-based one – by casting a vote. The mechanism that received the majority of votes was then applied to all participants in the respective group.

H&R employ a 2x2 experimental design, differentiating between real monetary and hypothetical incentives, as well as between easy and difficult tasks. While potential economic implications of overconfidence had been extensively discussed before, H&R were among the first experiments on overconfidence to incorporate monetary incentives. In the monetarily incentivized treatments, 50% of participants – in expectation – would receive a bonus payment. The allocation of this bonus was determined either by the performance in a vocabulary test (with the top 50% receiving the bonus) or by an individual die roll, where half of the possible outcomes (rolling a 4, 5, or 6) awarded the bonus[30]. Since real-life economic decisions are typically constituted by monetary stakes, the monetary incentive conditions should be considered the more relevant of the four treatments. As a second treatment manipulation, besides the nature of the incentive (real vs. hypothetical), task difficulty varied between groups. Test items were explicitly labeled as either easy or difficult and were designed accordingly.

An individual's voting choice served as a proxy for their confidence level relative to the group. Consequently, overplacement was inferred from a majority of participants preferring the performance test-based payoff, whereas a majority favoring the lottery-based mechanism was considered underplacement. H&R argue that voting for the performance-based mechanism is a weakly dominant strategy for anyone who believes they are more skilled than the median participant, assuming participants aim to maximize their personal payoffs. Therefore, the voting mechanism should ensure a truthful revelation of subjects' skill self-evaluation[31].

[30] In the treatments with hypothetical incentives, subjects were asked to imagine the respective payoff structure.

[31] For a more extensive discussion of the game's equilibria see 1.1.3 (p. 310) in H&R as well as their Technical Appendix

The experimental procedure was conducted using a seven-page questionnaire that was distributed and collected sequentially. Subjects were recruited from the university campus and participated in groups of at least seven. Before casting their vote on their preferred payoff mechanism, subjects were provided with two sample items and their respective solutions. Participants then completed both the knowledge test *LEWITE* ("Lexikon Wissenstest") [Wagner-Menghin, 2004][32] and the lottery task, without knowing which payoff scheme would ultimately be implemented until the experiment's conclusion. This design choice aimed to prevent biases in voting behavior arising from procedural preferences, such as effort avoidance[33]. Participants indicated their expectations for the test beforehand and reflected on it afterward through standardized Likert-scale questions. Additionally, they were asked to predict their own test performance in advance and estimate both their own and the group's scores after completing the test.

The experimental results notably do not support the presumption of a robust general human tendency for overconfidence in self-assessments, as voiced by DeBondt and Thaler [1995]. While H&R observe the majority of subjects favoring the performance-based bonus allocation scheme in three out of their four groups, the *Difficult x Money* treatment stands out from the others. Under real monetary incentives with hard test items, the majority vote shifts toward the luck-based payoff scheme (die roll), presumably because subjects perceive their chances of obtaining a bonus to be higher in a fair 50/50 lottery than in a difficult knowledge test. H&R infer overplacement tendencies for both easy and difficult tests under hypothetical incentives. Only when real monetary incentives are provided do individuals display a tendency for underplacement in relative self-assessments. Logit regression analysis of voting for the test on the expectation of performing better than the group average reveals an overall positive coefficient that becomes insignificant for the difficult test with money at stake. It is therefore concluded that a combination of monetary incentives and a difficult task discourages individuals from choosing a performance-based payoff scheme, even when they anticipate outperforming their competition.

5.3 Limitations to H&R and Derivation of Hypotheses

More recent literature on performance- vs. random device-based payoff scheme choices has hinted that the voting patterns observed in H&R may not necessarily indicate underconfidence but rather reflect aversion to risk or ambiguity associated with the test [Owens et al., 2014; Blavatskyy, 2009; Grieco

under A.3.

[32] This test featured a mechanism for comparing self-assessment with an immediate knowledge check. First, participants indicated which of the twenty listed words they knew and could explain. Second, they were presented with explanation sentences for all twenty words and were required to fill in two blanks for each definition by selecting from seven to nine answer options. The structure of the test provided an additional measure of inaccurate self-assessment beyond the vote and made success through guessing highly unlikely.

[33] For a general discussion of the effect of equal payments on effort see Abeler et al. [2010].

and Hogarth, 2009]. H&R themselves acknowledge that under economic incentives with a difficult test their measure may be skewed toward the lottery. While the lottery features a clearly defined win probability, the capabilities of one's competitors in the performance test – and the resulting probability of winning – are ambiguous. However, individuals' willingness to sacrifice potential earnings or pay an insurance premium to avoid ambiguity has been pointed out in the literature [see e.g., Ellsberg, 1961; Hogarth and Kunreuther, 1989; Fox and Tversky, 1995; Moore and Eckel, 2003], particularly when stakes are high [Blavatskyy and de Roos, 2023]. Meanwhile, ambiguity attitudes are typically confounded with risk preferences [Gneezy et al., 2015], which H&R abstract from in their equilibrium derivation of subjects' voting strategy. Consequently, the extent to which underconfidence drives the shift toward the lottery may be overestimated, with some of the observed voting behavior instead being explained by aversion against ambiguity and risk.

Beyond these limitations, which H&R mention in passing, little attention has been paid toward the structural differences of the test and lottery mechanisms. This study argues that the two payoff schemes are not fully equivalent, as their statistical properties allow for different distributions of bonus recipients while simultaneously exhibiting differing degrees of outcome-dependency. The performance test determines a fixed number of winners (with no variance: $\sigma^2 = 0$), which is known to participants ex ante given knowledge about group size, with individual win probabilities being interdependent. In contrast, an individual die roll provides each participant with an independent 50% chance of winning. Therefore, while the lottery assigns 50% winners in expectation, it introduces a variance to the actual number of winners ($\sigma^2 = n/4$), which, in the most extreme – though unrealistic – case, result in every single participant receiving a bonus.

The anticipation of this flexibility in the number of potential bonus recipients under the lottery scheme – in conjunction with the independence of win probabilities – may create a discrepancy in how the two mechanisms are perceived, particularly given the prospect of real monetary incentives and a difficult task. While the test condition constitutes a zero-sum environment – where a fixed number of bonuses are awarded and one participant's success necessarily precludes another's – the lottery does not impose such constraints. Thus, voting for the lottery option, which provides each group member with an independent and objective 50% chance of receiving the bonus, holds the potential to improving social welfare [Charness and Rabin, 2002] beyond what would be possible under the performance-based scheme. This reasoning aligns with findings from redistribution research, which suggests that "inequality is tolerated more if it does not come at the expense of others" [Strang and Schaube, 2025, p. 2].

Although a minority of individuals may be classified as purely selfish payoff-maximizers, most people exhibit at least some degree of altruism or fairness concerns and are often willing to forgo

personal monetary gains to either benefit others directly or reduce payoff inequalities [Andreoni and Miller, 2002; Loewenstein et al., 1989; Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999; Engelmann and Strobel, 2004; Fehr et al., 2006]. As a result, voting behavior may deviate from subjects' relative self-assessments, with some participants who could reasonably expect to place in the upper half of the group in terms of performance still opting for the lottery. This pattern should be less pronounced under hypothetical payment schemes or easier tasks with lower performance differentiation.

Conversely, inequality based on performance is often considered acceptable [Freyer and Günther, 2023; Gee et al., 2017], but primarily when the exerted effort is observable, as this allows people to distinguish between outcomes due to effort versus those due to luck or talent [Fischbacher et al., 2017], while people are more likely to attempt to mitigate inequality perceived to stem from luck [Mollerstrom et al., 2015; Rey-Biel et al., 2015]. This concern is arguably more relevant to the lottery mechanism – somewhat ironically, given that it is based on luck and could also result in fewer than 50% of participants winning – as it allows for variance in the number of bonus recipients and may thereby reduce inequality, unlike the test condition. Krawczyk [2012] notes that when bonus payments are conditional on success in a task or lottery, some individuals may prefer a test-based bonus allocation because they consider performance-dependent payment fair, while others may favor a lottery for its ability to reduce outcome inequality.

Accordingly, votes for the lottery-based payoff mechanism may reflect a broader and more diverse array of motives – particularly non-monetary ones – beyond purely reflecting relative self-assessment under payoff-maximizing constraints outlined in H&R's equilibrium strategy derivation. Their identification strategy relies on the assumption that subjects have an incentive to truthfully reveal their self-assessment relative to the group through their vote, as it goes along with the highest chance of receiving the bonus. Consequently, H&R's results likely overestimate participants' actual self-assessment-induced preference for the lottery-based payoff mechanism and, in turn, the extent of underconfidence.

Implementing an alternative lottery mechanism that converts the die roll into a random device with a fixed-outcome distribution and interdependent success probabilities should objectively equalize potential perceived disparities in the numbers of winners between the performance test and the lottery[34]. This adjustment may lead to voting results that more accurately capture individuals' subjective relative self-assessment, aligning more closely with the study's original intent.

Thus, the following hypotheses are proposed for comparing H&R's original experiment with the modified lottery mechanism:

[34]The modified lottery mechanism however does not affect the ambiguity associated with the test.

Hypothesis 1: Voting behavior differs between a lottery with an independent win probability of 50% and a lottery with a fixed-outcome distribution win rate of 50%.

Hypothesis 2: Specifically, less underplacement occurs with a fixed-outcome distribution lottery compared to the probabilistic lottery used in H&R.

5.4 Current Experiment

For this replication, conducted in late 2024 via the academic crowd-working platform Prolific[35], H&R's experiment was translated into an online format using oTree [Chen et al., 2016]. The study was pre-registered in the American Economic Association's registry for randomized controlled trials and was approved by the German Association for Experimental Economic Research e.V. (GfeW)[36] prior to the experiment. The replication focuses solely on the *Difficult x Money* group from H&R, which includes real monetary incentives and a difficult task, as it is the only condition in which behavior significantly differed from the other groups, while also being the most relevant from an economic perspective. The full experimental instructions are available in Appendix D.3, while a detailed comparison by aspect between this experiment and H&R's original version can be found in Appendix D.2.

5.4.1 Task

Using the LEWITE performance test as H&R was deemed impractical in an online setting, as participants could easily look up vocabulary definitions online. This issue would likely persist in a laboratory setting due to the accessibility of mobile devices. Therefore, a set of multiple-choice verbal analogy questions – previously used in the SAT analogy test [Turney et al., 2003] – was adopted instead[37]. This task, paired with a per-item time limit, was selected because it met multiple criteria for comparability with the original study. First, identifying analogies targets similar cognitive abilities as defining vocabulary, combining semantic, verbal, and logical reasoning skills. Second, analogies are not easily searchable online, especially under time pressure. Third, SAT questions feature a sufficient number of distractors (four) besides the correct response to reduce the likelihood of success through guessing, albeit not as effectively as LEWITE[38]. Finally, these analogies require advanced vocabulary and have been shown to be relatively complex [Church, 2017]. Therefore, they should be sufficiently challenging

[35]Prolific was chosen as it allows for the application of specific sampling criteria [Palan and Schitter, 2018], and its participant pool has been shown to be more attentive than those on other platforms such as MTurk [Albert and Smilek, 2023]. For a more detailed comparisons with other platforms, see Peer et al. [2017].

[36]AEA RCT Registry in May 2024: https://doi.org/10.1257/rct.13070-1.0.
GfeW Registry in February 2024: https://gfew.de/ethik/4dqGpmS5

[37]Example question: Which pair of words relates to each other like "liter" to "volume"? A) day - night; B) mile - distance; C) decade - century; D) friction - heat E) part - whole. Answer "B)" is correct here because the first word represents a unit of measurement for the second as "liter" does for "volume".

[38]The LEWITE scheme consists of a sentence with two gaps and seven to nine answer options. In the easiest case, guessing would therefore be successful in $1/7 * 1/6 = 1/42 = 2.38\%$ of attempts, assuming independence between answers for simplicity). In comparison, the success probability of guessing on an SAT item is $1/5 = 20\%$.

– particularly in combination with a per-item time limit – to be considered "difficult" in the sense of H&R. Also, they should allow for enough performance variance to create an informative ranking for the test-based payoff condition. Like in H&R, the performance test consisted of 20 items. From a pool of 337 non-validation SAT test items, 23 were randomly selected, translated to German, and checked for potential region-specific terms. Twenty items were then used as tasks for the performance test, while the remaining three served as sample questions (see Appendix D.3).

Potential limitations of the alternative task format will be discussed in Section 5.6.

5.4.2 Sample

A between-subject experimental design was used to compare two experimental groups of one-hundred participants each, one using a probabilistic-outcome distribution lottery mechanism mimicking the one from H&R and one incorporating a modified fixed-outcome distribution version (see 5.4.4 and 5.4.5). The sample size was determined via power analysis using G* Power 3.1 [Faul et al., 2009]. Following the experimental design with the goal of comparing two independent groups, assuming a medium effect size $d = 0.4$ [Cohen, 2013], and using the social science convention of $a = 0.05$ (two-tailed), 80% power [Brysbaert, 2019], 100 subjects per group were needed for sufficiently powered results, resulting in a total sample size of two-hundred subjects ($N = 200$).

At the time of the study, the Prolific participant pool contained 4,475 individuals who matched the study specification of "first language: German". This criterion was deliberately selected to align with the original language of H&R's study and its subject pool. Each subject was allowed to participate in only one experimental condition. While H&R's sample was mainly comprised of student participants recruited from the university campus, the Prolific worker base is more heterogeneous in terms of socio-demographic characteristics 45). Prolific is a well-established platform for recruiting participants for academic studies from whom a certain level of knowledge, cognitive abilities, and effort can be expected. Additionally, testing the robustness of H&R's results to a non-student sample should be insightful in itself. Unlike H&R's original study, which conducted multiple smaller sessions, this experiment featured a single large group per treatment condition, altering the number of individuals against whom subjects form their relative self-assessment. The potential implications of this aspect are discussed in Section 5.6.

5.4.3 Procedure

An overview of the basic experimental procedure is provided in Figure 17.

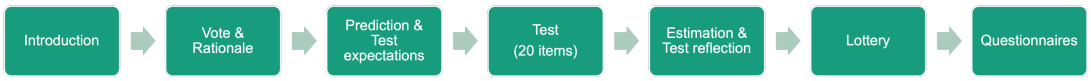


Figure 17: Overview of experimental procedure

After being thanked in advance for their participation and providing consent for data processing, participants received the experimental instructions. A countdown of 90 seconds was set as minimum reading time before they could proceed to the next page. The instructions provided detailed information about the voting decision, the two bonus payment mechanism alternatives, and the subsequent experimental procedure. Next, extensive comprehension checks on key aspects of the experimental design were conducted to ensure proper understanding of the rules. Participants were only allowed to advance to the vote after correctly answering all comprehension checks.

Similar to H&R, subjects were then presented with three sample performance test items, which they could solve to familiarize themselves with the task type. As in H&R, participants were informed that the test items could be categorized as difficult. On the subsequent page, the correct solution to the sample items were displayed alongside the participants' answers. Subjects were then informed that the vote would take place on the following page.

To cast their vote, participants simply clicked a radio button next to either "Test" or "Lottery". The explanations of the two payoff mechanisms were displayed above the selection buttons to ensure proper understanding. Beyond H&R's procedure, after casting their vote, participants were asked to provide a brief rationale (at least 20 characters) for their choice to better understand their voting motives and gain additional insight into their decision-making process.

Next, exactly as in H&R, subjects predicted their own score as well as the average score among all participants before proceeding to the performance test. Furthermore, they were asked how sure they were to have made the right choice in the vote, how difficult they would find it to change their decision, how important was doing well in the test to them, how difficult they thought the test would be and how well they thought they will do in it.

Unlike in H&R, the test included a 30-second countdown per item to prevent participants from searching for answers online – an issue that was not relevant at the time of the original study.

After completing all 20 test items, participants were shown their test score along with the correct solutions. They were subsequently asked to (re-)estimate the average test score among all participants and to indicate their satisfaction with their test results, how difficult they found the test, how sure they were now to have made the right choice in the vote and how difficult they would find it now to change it (all questions identical to those in H&R). Subsequently, the lottery part commenced with participants selecting a "lucky number" between 1 and 6. The lottery design constituted the

experimental treatment variation (described in detail in Sections 5.4.4 and 5.4.5).

Following the lottery, participants completed a multi-part questionnaire that included three standardized scales, two price list tasks, and basic demographic questions.

Since social comparison aversion and low self-efficacy may be potential reasons discouraging individuals from voting for a performance-based payoff scheme test in favor of the lottery, the questionnaire contained a six-item short version of the Iowa–Netherlands Comparison Orientation Measure (INCOM) [Schneider and Schupp, 2014; Gibbons and Buunk, 1999] to assess participants' inclination toward engaging in social comparison, as well as short-measure for general self-efficacy [Beierlein et al., 2013], serving as controls. Also, the altruism sub-scale of the HEXACO-PI-R-100 [Lee, 2018] was included as a measure of individuals' prosocial tendencies. Afterward, participants completed the multiple price list format by Dohmen et al. [2011] to elicit individual risk preferences as well as the analogously designed lottery task by Gneezy et al. [2015] to measure individual attitudes towards ambiguity[39]. While ambiguity aversion usually appears stable for same individual on different tasks, related literature emphasizes the importance of jointly assessing risk and ambiguity attitudes, as otherwise either may be overestimated [Krahnen et al., 2014; Gneezy et al., 2015]. Additionally, the simple self-report measure of general risk preferences by Dohmen et al. [2011] was included as a control for response consistency since it is more intuitive than the lottery format. Finally, subjects answered demographic questions regarding their age, gender, education level, student status and study major or professional occupation.

To ensure attentiveness, two attention checks were embedded in the experiment – one at the very beginning and another within the INCOM questionnaire. At the end of the study, participants were also asked directly whether they had paid attention and completed the experiment in one go. Their response had no effect on their compensation, allowing for voluntary disclosure of inattentiveness or carelessness.

5.4.4 Replication Condition

In the Replication condition, the analogous die roll from H&R was emulated by a random generator draw. Subjects were asked to choose a "lucky number" between 1 to 6. After all 100 subjects had participated, a random number between 1 and 6 was drawn. If this number was even, all subjects who had chosen an even lucky number received the bonus payment, while those who had chosen an odd number won if an odd number was drawn.

Thus, subjects had a 50% chance of receiving a bonus payment, independent of other subjects'

[39]Neither task was incentivized to prevent dilution or hedging of incentives from the main experiment. Prior research has shown that response behavior in these types of lotteries does not significantly differ between medium-to-high hypothetical incentives and small real incentives [Holt and Laury, 2002; Gneezy et al., 2015].

choices, in contrast to the test condition, where an individual's chance of winning partly depended on the performance of others. Consequently, this yielded a probabilistic-outcome distribution of 50 winners and 50 non-winner in expectation, meaning that among all 100 participants, 50% would, on average, receive the bonus payment (provided the majority vote favored the lottery payoff mechanism). However, the actual outcome distribution would deviate from an exact 50/50 split if subjects' number choices were non-uniformly distributed, just as could happen with outcomes in a traditional die roll.

5.4.5 Adaptation Condition

In the Adaptation condition, the lottery format was modified to ensure a fixed-outcome distribution, eliminating variance in the number of bonus recipients. Specifically, the lottery determined exactly 50 winners and 50 non-winners through random draw. As a result, the lottery-based payoff scheme becomes more comparable to the performance-based one, since – unlike in the Replication condition – the number of winners is predetermined, with no variance. Also, similar to the test condition, the chance for receiving a bonus payment becomes interdependent among subjects. Consequently, instead of being binomially distributed as in the Replication condition, the probability of receiving a bonus follows a hypergeometric distribution.

To maintain procedural parallelism with the Replication condition, and to avoid considerations related to a potential illusion of control [Langer, 1975; Presson and Benassi, 1996], subjects in the Adaptation group were also asked to choose a "lucky number" from 1 to 6. Together with a subject's Prolific worker ID, the chosen number formed their unique "win code". After all 100 subjects had participated, 50 codes were drawn at random to determine which subject would receive a bonus payment (provided the majority vote favored the lottery payoff mechanism). If the majority vote favored the performance-based bonus payoff, conditions remained exactly the same as in the Replication condition.

5.4.6 Incentive Scheme

Subjects received a fixed payment of GBP 3.00 and were informed about the conditions for receiving a bonus payment in the instructions. The possibility of receiving a bonus payment was also prominently announced on the study's invitation page. The experiment was estimated to last approximately 25 minutes, based on test runs with student assistants, while H&R's original experiment took around 30 minutes. This translates to an hourly rate GBP 7.20, which is categorized as "fair" by Prolific (the minimum required hourly reward is GBP 6.00). However, since – in expectation – half of all subjects would receive the bonus payment, the average hourly earnings increased to GBP 13.20, exceeding Prolific's recommended rate of GBP 9.00 and making it competitive compared to other economic

experiments.

Once all 100 subjects in a group had completed the experiment, a bulk mail was sent via the Prolific platform, announcing the the voting result and the median test score. The median test score automatically determined the cutoff for bonus payments under the performance-based scheme. Alternatively, if the lottery-based scheme was chosen, the worker IDs of those subjects receiving the bonus were listed. The bonus payments were credited to the respective participants' accounts shortly afterward.

5.5 Results

A total of three-hundred individuals ($N = 300$) participated in the experiment, completing it in an average of 23 minutes. Originally, the research plan aimed for 100 participants per treatment condition, totaling 200 participants. However, the two original sub-samples exhibited a statistically significant gender distribution imbalance. Coupled with an observed gender difference in voting tendencies – 58.9% of female participants voted for the test, compared to 75.2% of male participants (Pearson $\chi^2(1) = 5.42, p = 0.020$) – this created inference issues in assessing a potential treatment effect. To control for potential gender bias and ensure an unbiased inference of treatment effects without confounding factors, a second, gender-balanced sample of 100 participants was drawn for the Replication condition[40].

Although all participants affirmed to have paid sufficient attention and effort, 18 observations were excluded from the analysis: 10 from the Replication group and 8 from the Adaptation group. Nine failed at least one attention check, eight provided voting rationales that were nonsensical, arbitrary or inconsistent to their vote, and one explicitly stated an intent to bias the results. Thus, 182 subjects enter the subsequent analysis. The resulting sample had an average age of 35.9 years (Std. dev. = 11.8; median = 33.0), ranging from 19 to 74. This was noticeably higher than in H&R (23 years), reflecting the broader demographic composition of Prolific compared to H&R's predominantly student-based sample. The sample was 51% female. The most common fields of study and professions were "Business, Economics, Marketing, Sales, or Insurance' and 'Education, Cultural Studies, or Public Sector,' each representing close to 20% of participants, followed by 'Information Technology and Natural Sciences' at just over 15%. Appendix D.2 provides a detailed demographic overview.

5.5.1 Choice of Vote

Unlike in H&R, the test payoff mechanism received the majority vote in both treatments, with 67 and 66 votes in the Replication and Adaptation condition, respectively. As shown by one-sided Binomial

[40]The original Replication group comprised 30% women and 70% men, whereas, by chance, the gender distribution in the Adaptation group was exactly balanced at 50/50 (Pearson $\chi^2(1) = 7.56, p = 0.006$). Since Prolific offers the possibility to recruit samples with predefined demographic criteria, this feature was leveraged to enhance the comparability and of the results.

tests in Table 19, the voting share in favor of the test significantly exceeds the 50% reference point in both groups.

Table 19: Average Vote for Test, by Treatment

	Mean	Std. dev.	Binomial test p-value
Replication	0.744	0.439	0.0000
Adaptation	0.717	0.453	0.0000

However, the distribution of voting shares does not differ significantly between the Replication and Adaptation groups (Pearson $\chi^2(1) = 0.17, p = 0.681$). Accordingly, the data do not support Hypothesis 1.

Whereas H&R observed underplacement, the voting distributions in both the Adaptation and Replication conditions suggest overplacement. Thus, while the test received significantly more votes than the 39% observed in H&R ($p < 0.0001$ for one-sided Binomial tests in both groups) which means 'less underplacement' than in H&R in a literal sense, this difference is clearly not attributable to the voting mechanisms' characteristics. Consequently, Hypothesis 2 is also not supported.

In the initial gender-unbalanced Replication sample, 65.5% of participants voted for the test – based on 90 observations that remained in the sample after applying the same exclusion criteria as in the other groups – which would also have led to the rejection of the hypotheses (Pearson $chi2(1) = 0.8087, p = 0.369$). Since the lottery mechanism's design does not appear to alter voting outcomes between treatments, all available observations ($N = 272$) are pooled for the subsequent analysis in order to examine the factors and motivations underlying individual voting behavior. Additional treatment comparisons are provided in Table 46 in Appendix D.2.

5.5.2 Subjects' Assessments

On average, subjects predicted their own performance to be nearly 14 correct answers out of 20, which is approximately two more than the average in H&R's incentivized treatments (11.35). Correspondingly, their own performance prediction exceeded their prediction of the group's mean performance by nearly one point (see Table 20), indicating a statistically significant difference in expected performances (Wilcoxon signed-rank test: $|z| = 5.64, p < 0.0001$). Meanwhile, subjects' predictions of the group's mean performance closely aligned with the corresponding prediction in H&R (13.09). Using the ratio between the two predictions – H&R's "better" indicator – subjects, on average, expected their own performance to be 8.8% above the group mean. This contrasts with the direction of H&R's findings, where subjects in the *Money × Difficult* treatment reported a ratio of 0.88, indicating they expected to perform, on average, 12% worse than their peers.

Table 20: Subjects Performance Predictions and Ratio Before the Test

	Pooled sample	
Variable	Mean	Std. dev.
Predicted own performance	13.94	3.612
Predicted group performance	12.99	2.594
Better	1.088	0.269

Overall, nearly two-thirds (63.97%) of participants predicted outperforming the group average, which is a proportion significantly higher than 50% (One-sided Binomial test: $p = 0.0000$) and substantially higher than in H&R, where only 33% of subjects made such predictions, peaking at 47% in the *Easy* × *No Money* treatment.

Examining the consistency between performance predictions and voting behavior (see Table 21), 87.4% of subjects who predicted to outperform the group voted for the performance-dependent payoff mechanism (i.e., the test), compared to 40.8% of those who predicted performing at or below the group average. This substantial and statistically significant difference (McNemar's $\chi^2(1) = 5.23, p = 0.0300$) underscores the connection between self-assessment and choice of compensation scheme. Mathematically equivalent, 79.6% of test voters expected to outperform the group average, compared to only 27.5% of lottery voters.

Table 21: Consistency of Predictions and Voting Behavior

Prediction	Test	Lottery	Total
Own > Group	152	22	174
Own ≤ Group	40	58	98
Total	192	80	272

Accordingly, while the majority subjects' votes aligns with their self-assessments, notable deviations from equilibrium strategy behavior persist. From an equilibrium perspective, *all* subjects expecting outperform the group should have chosen the test; yet, over 12% opted for the lottery. More strikingly, more than 40% of those anticipating their performance to be at or below the group mean still voted for the test. These discrepancies suggest that factors beyond self-assessment likely influenced voting behavior (see Section 5.5.4).

On average, participants solved 12.8 correctly (see Table 22), falling short of their own predicted performance by 1.1 points. This gap is notably larger than in H&R's *Difficult* treatments (approx. 7 correct answers) but smaller than in their Easy treatments (approx. 16 correct).

Table 22: Subject Performance and Estimated Group Performance After the Test

	Pooled sample	
Variable	Mean	Std. dev.
Actual own performance	12.84	3.551
Estimated group performance	12.52	2.594

Subjects' own performance predictions were significantly correlated with their actual performance (Spearman's $\rho = 0.2435, p = 0.0001$), indicating that individuals were quite good at assessing their performance, in tendency. As in H&R, in a simple linear regression of actual performance on predicted performance yielded a positive and significant coefficient, though its magnitude was roughly one-third of that in H&R (see Table 47 in Appendix D.2). The coefficient of determination being small was attributed to large individual errors by H&R. A similar pattern can also be observed in this study, with subjects' estimates of their own scores differing significantly from their actual performance (Wilcoxon signed-rank test: $|z| = 4.01, p = 0.0001$). The discrepancy featured substantial variation (Std. dev. = 4.5 points) caused by instances of extreme overestimation (up to 16 points) and underestimation (up to 10 points), as can be seen in Figure 18. Nevertheless, 30 participants (11%) precisely predicted their own score, while one-third (33.1%) of provided predictions within one point of their actual score. Overall, a tendency for overestimation can be observed, with more than half of subjects' (55.5%) own performance predictions exceeding their actual performance. Simultaneously, one third (33.5%) of subjects underestimated their own performance.

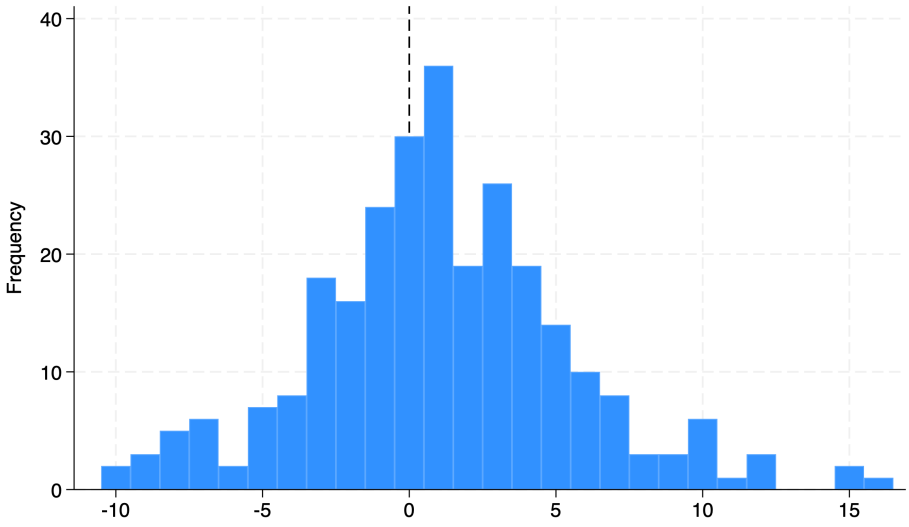


Figure 18: Frequency distribution of differences between predicted and actual performance

Following test completion and learning their individual scores, participants revised their estimates of group performance downward by an average of 0.5 points – a statistically significant difference from

pre-test predictions (Wilcoxon signed-rank test: $|z| = 2.57, p = 0.0100$). Despite this adjustment, the difference between subjects' own performance (now known with certainty) and their estimated group mean remained significant (Wilcoxon signed-rank test: $|z| = 2.87, p = 0.0041$), though the absolute gap narrowed. This pattern aligns with Moore and Healy [2008, p. 511]'s prediction that individuals' "beliefs about their own scores should be a joint function of their prior expectations and the private signals they receive regarding their own performance (i.e., their experience taking the quiz)". The observed downward revision suggests that subjects extrapolated their own underperformance onto their group estimate while maintaining the directional belief that they outperformed the group, albeit with a reduced margin. This adjustment may be explained by subjects placing greater weight on their own performance signal than on beliefs about peer performance, as noted by [Moore and Cain, 2007]. Meanwhile, the proportion of participants estimating the group to have performed worse or equally to themselves (54.8%) shifts significantly closer to those estimating the group to have performed better (45.2%), compared to pre-test predictions (McNemar's $\chi^2(1) = 7.53, p = 0.0080$).

Item-wise comparisons of subjects' responses to pre- and post-test perception questions – such as being sure about one's voting decision, importance of doing well in the test, expected task difficulty, and expectation of doing good – between this study and H&R can be found in Tables 48 and 49 in Appendix D.2. In the pre-test assessments, scores for all five items are higher in this study. While the level of certainty about having made the right voting decision ("sure") is only marginally higher than in H&R, the differences in the other four measures are substantial and highly statistically significant. Most notably, perceived task difficulty and the stated importance of doing well appear particularly elevated, exceeding H&R's values by 0.9 and 2.7 points, respectively, on a 7-point scale. However, these differences must be interpreted with caution, as H&R only report averages across their entire sample despite the likelihood that these expectations were influenced by at least one of their treatment manipulations. For instance, the actual task difficulty should affect both perceived task difficulty and subjects' performance expectations[41]. Similarly, while real monetary incentives – which apply to all participants in this study – should affect importance placed on performing well and, potentially, both subjects' certainty level about their voting decision and the perceived ease to change it. The same considerations apply to post-task reflections.

After the test, subjects in this study reported slightly lower satisfaction with their performance compared to those in H&R – potentially due to performing well in the sample task but falling short in the actual test. Additionally, they were significantly less sure about having made the correct voting

[41] Perceived task difficulty before the test was significantly correlated to subjects' expectation to do good in it (Spearman's $\rho = -0.3659, p < 0.0001$). Meanwhile, the expectation to do good was significantly correlated with their predicted score (Spearman's $\rho = 0.6296, p < 0.0001$) and, consequently, with the prediction ratio "better" (Spearman's $\rho = 0.5105, p < 0.0001$), but only weakly and insignificantly correlated with their actual performance (Spearman's $\rho = 0.1062, p = 0.0803$).

decision post-test than both subjects in H&R (Two-sample t-tests, equal variances: $t = -2.47, p = 0.0139$) and themselves pre-test (Wilcoxon signed-rank test: $|z| = 3.11, p = 0.0018$), with a decline of approximately 0.4 points in both cases. This shift can likely be attributed to the signal received through learning their test results (see Sections 5.5.3 and 5.5.4), as quiz scores and post-test certainty level are significantly correlated (Spearman's $\rho = 0.3162, p = 0.0000$)[42]. Furthermore, perceived task difficulty increased slightly (by 0.2 points on average) but significantly (Wilcoxon signed-rank test: $|z| = 3.35, p = 0.0008$) from pre-test expectations to post-test assessments, indicating that subjects found the task somewhat more challenging than anticipated. This pattern contrasts with H&R, where perceived difficulty did not increase post-test. In absolute terms, subjects in H&R rated their task as 1.2 points easier than those in this study. Notably, the fact that subjects, on average, still estimated their own performance to be superior to that of their peers despite perceiving the task as more difficult than expected contrasts with the findings of Healy and Moore [2007]. Meanwhile, subjects' perceived difficulty in changing their vote remained virtually unchanged from pre- to post-task, whereas in H&R, it increased by half a point. However, it is important to note that in H&R, the die roll took place immediately after the test with the winning numbers known at the time. As a result, subjects were aware of whether they would receive a bonus if the lottery won the vote when completing the post-test questionnaire. This additional information may have influenced their assessment of both their level of being sure about their vote and their willingness to change it. In contrast, in this study, the lottery outcome was revealed only after all subjects had participated. Lacking this key information may explain why subjects' certainty level about the voting decision was lower than in H&R.

5.5.3 Predictors and Correlates of vote

Following the analysis from H&R, a logit regression of the vote on "better" – defined as the ratio of predicted own performance to predicted group performance, serving as implicit measure of self-assessment relative to the group – reveals a positive relationship between the probability of voting for the test and assessing one's own performance as superior to the group's average (see Table 23). This relationship is highly significant, with a marginal effect (at the mean) of 0.769 ($p < 0.0001$), indicating a 76.9 percentage-point increase in the likelihood of voting for the test if a a predicts their own performance to be superior to the group's average.

[42]This correlation is clearly driven by test voters, as the relationship is particularly strong and highly significant among them (Spearman's $\rho = 0.5094, p < 0.0001$), whereas no such correlation is observed among lottery voters (Spearman's $\rho = 0.0071, p = 0.9499$). This pattern is intuitive: subjects who voted for the test and learned they performed well could – in a vacuum – conclude to have made a good decision, despite having no information on their relative standing yet.

Table 23: Logit regression of Vote for Test Over Ratio of Performance Predictions

	pseudo-R^2	χ^2	Prob $> \chi^2$
	0.2209	37.61	0.0000
Vote for Test	Coefficient	z	$P > z$
Better	5.0265	6.13	0.000
Constant	-4.3622	-5.23	0.000

Note: *Better*: Individual ratio between predicted own performance and predicted group. Regression model estimated using robust standard errors. n = 272.

Consistent with H&R, a logit regression of test voting behavior using predicted own performance and predicted group performance as independent variables yields highly significant coefficients of the expected signs (see Table 24). Predicted own performance positively affects the likelihood of voting for the test (marginal effect at the mean: $0.066, p < 0.0001$), whereas predicted group performance has a negative effect (marginal effect at the mean: $-0.050, p < 0.0001$). Thus, a one-point increase in predicted own performance is associated with a 6.6 percentage-point increase in the likelihood of voting for the test, while a one-point increase in predicted group performance corresponds to a 5.0 percentage-point decrease.

Table 24: Logit regression of Vote for Test Over Performance Predictions

	pseudo-R^2	χ^2	Prob $> \chi^2$
	0.2500	45.23	0.0000
Vote for Test	Coefficient	z	$P > z$
Predicted own performance	0.4550	6.72	0.000
Predicted group performance	-0.3472	-4.40	0.000
Constant	-0.7580	-0.86	0.390

Note: Regression model estimated using robust standard errors. n = 272.

Signals Furthermore, it appears reasonable to assume that subjects incorporated information from solving the sample questions into their self-assessment and voting decision. Additionally, this information may have influenced their assessment of the group's performance, despite the absence of direct feedback on competitors' performance.

On average, subjects correctly answered 2.28 out of 3 sample questions (Std. dev. = 0.843). However, as they had unlimited time and no restrictions against external assistance, cheating cannot be ruled out. In fact, sample question performance is strongly and significantly correlated with performance in the main task (Spearman's $\rho = 0.498, p < 0.0001$). Additionally, higher sample question performance is associated with with higher own performance predictions (Spearman's $\rho = 0.417, p < 0.0001$)[43]. In line with these correlations, subjects who performed better on the sample questions

[43]Similarly, a significant correlation of sample question performance with both the pre-task control question on how good subjects think they will do in the test (Spearman's $\rho = 0.209, p = 0.0005$) and the ratio of own and group predictions

were significantly more likely to vote for the test as the payoff mechanism. The proportion of test voters increased progressively, from 10.0% among those who solved no sample questions correctly, to 43.6% with one correct, 70.5% with two correct, and 83.0% with all three correct responses (Pearson $\chi^2(3) = 41.33, p < 0.0001$)[44]. Similarly, when extending the logit regression of voting behavior on the two predictions by adding sample performance as a predictor (see Table 25), the marginal effect (at the mean) of sample performance on test voting probability is 0.105 ($p < 0.0001$). This inclusion reduces the marginal effect of own performance prediction while increasing the overall explanatory power of the model.

Table 25: Logit regression of Vote for Test Over Performance Predictions and Sample Performance

	pseudo-R^2	χ^2	Prob $> \chi^2$
	0.2916	47.72	0.0000
Vote for Test	Coefficient	z	$P > z$
Predicted own performance	0.3996	5.73	0.000
Predicted group performance	-0.3825	-4.55	0.000
Sample question performance	0.7745	3.35	0.001
Constant	-1.2589	-1.39	0.164

Note: Regression model estimated using robust standard errors. n = 272.

Interestingly, higher sample question performance was also associated with increased group performance predictions (Spearman's $\rho = 0.2685, p < 0.0001$). This finding somewhat contradicts the argument by H&R that individuals fail to adjust their expectations of others' performance to newly acquired task-related information.

As it could be expected, individuals who voted for the test reported a higher confidence higher expectation to do well in it, averaging 5.2 on a 7-point scale, compared to 4.3 among lottery voters (Two-sample Mann-Whitney U-Test: $|z| = 5.35, p < 0.0001$). Test voters also assigned greater importance to performing well than lottery voters (6.1 vs. 5.6 on a 7-point scale), though the latter still reported relatively high importance (Two-sample Mann-Whitney U-Test: $|z| = 2.69, p = 0.0073$). Perceived task difficulty did not differ significantly before the test (Two-sample Mann–Whitney U-test: $|z| = 1.73, p = 0.0832$). However, lottery voters perceived the task as slightly more difficult beforehand and downgraded their assessment afterward (5.6 to 5.4, Wilcoxon signed-rank test: $|z| = 1.17, p = 0.2434$), while test voters initially perceived the task as less difficult but upgraded their assessment the task and upgraded perceived difficulty (5.4 to 5.8, Wilcoxon signed-rank test: $|z| = 4.67, p = 0.0000$). The level of being sure about their voting decision and the perceived difficulty to change did not differ by vote.

("better") can be observed (Spearman's $\rho = 0.184, p = 0.0024$). Moreover, sample question performance is negatively correlated with perceived task difficulty before the test (Spearman's $\rho = -0.1826, p = 0.0025$).

[44]The share of test voters increases significantly from each number of correctly solved sample questions to the next.

Questionnaire Controls Table 26 presents the mean scores for the additional questionnaire controls introduced in this study (beyond those in H&R) across the full sample, as well as differentiated by voting decision.

Table 26: Questionnaire Summary Statistics

			Vote			
	Pooled sample		Test		Lottery	
Variable	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
INCOM	3.47	0.74	3.44	0.73	3.55	0.78
GSE	3.87	0.68	3.82	0.72	3.89	0.66
Altruism	3.79	0.69	3.77	0.67	3.84	0.74
Risk (MPL)	15.26	5.72	15.19	5.47	15.45	6.31
Risk (11-point)	5.51	2.47	5.43	2.50	5.69	2.42
Ambiguity	12.24	5.54	12.47	5.61	11.70	5.35

Note: INCOM: Short scale of Iowa–Netherlands Comparison Orientation Measure (5-point scale); *GSE*: General self-efficacy (5-point scale); *Altruism*: Altruism facet from HEXACO-100 (5-point scale); *Risk (MPL)*: Multiple price list format for measuring risk preferences; *Risk (11-point)*: General willingness to take risk (11-point scale); *Ambiguity*: Multiple price list format for measuring ambiguity aversion; N = 272; Test n = 192; Lottery n = 80.

Participants exhibited a moderate tendency for social comparison, averaging 3.5 on a 5-point scale. While one might intuitively expect individuals with a stronger inclination for social comparison to favor the test, the reported value is slightly – but insignificantly-higher among lottery voters (Two-sample Mann–Whitney U-test: $|z| = 0.95, p = 0.3448$). The same pattern holds for general self-efficacy. Although medium to high levels of self-efficacy were observed across the sample, the expectation that individuals with higher self-efficacy would be more inclined to enter a performance-based competition is not reflected in the data (Two-sample Mann–Whitney U-test: $|z| = 0.94, p = 0.3509$). Similarly, participants demonstrated medium to strong levels of altruism, yet no statistical difference was found between test and lottery voters (Two-sample Mann-Whitney U-test: $|z| = 0.73, p = 0.4684$), providing no evidence that altruism drives the decision to vote for the lottery. H&R themselves acknowledged ambiguity aversion as a potential influence on voting behavior, given that participants were choosing between a lottery with known probabilities and a test with an ambiguous probability of success. Using the elicitation price list from Gneezy et al. [2015], the average switching point from a lottery with known probabilities to ambiguous lottery occurred between rows twelve and thirteen, suggesting a general tendency toward ambiguity aversion, with participants demanding an ambiguity premium of approximately one-fourth of the potential winning amount. However, switching behavior did not differ significantly between test and lottery voters (Two-sample Mann-Whitney U-test: $|z| = 1.03, p = 0.3039$), and the distributions of switching points[45] were virtually identical (Two-sample Kolmogorov–Smirnov test: $D = 0.1000, p = 0.587$). Self-reported risk attitudes – measured

[45]To circumvent multiple switches between columns, the procedure suggested by Andersen et al. [2006] has been adopted in which only one row has to be designated for a switch from the lottery to the certainty equivalent.

using both an 11-point general risk scale and a multiple price list format [Dohmen et al., 2011] – located, on average, around the midpoint of the respective scales. Participants' willingness to take risks was relatively consistent between the two measures (Spearman's $\rho = 0.289, p < 0.0001$), and may overall be classified as approximately risk-neutral, though both measures exhibited standard deviations of around 40%. Neither measure revealed significant differences between test and lottery voters (Two-sample Mann-Whitney U-tests: General risk question: $|z| = 0.61, p = 0.5424$; Multiple price list: $|z| = 0.52, p = 0.6013$). Meanwhile, participants' risk and ambiguity attitudes were significantly correlated (Spearman's $\rho = 0.135, p = 0.0257$).

In summary, contrary to initial expectations, none of the additional latent constructs measured – namely social comparison tendency, general self-efficacy, altruism, risk preferences, or ambiguity attitudes – differed significantly between test and lottery voters (nor could any treatment differences be observed; see Table 46 in Appendix D.2).

Regarding demographic controls, no significant age differences were found between test and lottery voters (Two-sample Mann-Whitney U-test: $|z| = 0.66, p = 0.5101$). Meanwhile, as previously noted, voting behavior appears to differ significantly by gender: while 62.5% of female subjects voted for the test, 75.2% of male subjects did (Pearson $\chi^2(1) = 6.77, p = 0.009$). Additionally, 78.4% of students voted for the test compared to 66.8% of non-students, a borderline statistically significant difference (Pearson $\chi^2(1) = 3.83, p = 0.050$). This discrepancy could be attributed to students ascribing more intelligence to themselves or being more accustomed to test-taking, despite lacking information about the cognitive abilities and student status of other participants. Moreover, a marginally insignificant difference in voting behavior was observed between subjects with a degree in higher education (i.e., current or former students) and those without one: 74.9% of those with a degree voted for the test, compared to 63.8% of those without (Pearson $\chi^2(1) = 3.79, p = 0.052$). Interestingly, actual test performance did not differ significantly between men and women (Two-sample Mann-Whitney U-test: $|z| = 0.22, p = 0.8247$) or between students and non-students (Two-sample Mann-Whitney U-test: $|z| = 1.11, p = 0.2671$), which would justify the preference of the former for the test[46].

For completeness, Appendix D.2 presents the results of a logistic regression model (Tables 50 and 51) examining the likelihood of voting for the test, based on all previously discussed variables – including performance predictions, sample question performance, demographics, pre-test assessments, and questionnaire controls. These results largely support and reinforce the findings from non-parametric analysis. Performance predictions and signals from sample items remain strong and highly significant predictors of voting the test across all model specifications, while gender and student status also appear to be relevant factors, as discussed. A small yet significant negative effect of social comparison

[46]Similarly, no significant differences were found between these groups in sample question performance (Two-sample Mann–Whitney U-tests: Gender: $|z| = 1.70, p = 0.0896$; Student: $|z| = 0.49, p = 0.6352$).

orientation emerges, indicating that a stronger tendency to engage in social comparison is associated with a lower likelihood of voting for the test. Contrary to initial expectations, this may suggest that, at least in this context, participants who are more inclined to compare themselves to others tend avoid competition, possibly influenced by the perceived difficulty of the task. However, this interpretation remains speculative.

Calibration According to the equilibrium strategy outlined by H&R, a rational decision maker should vote for the test if they assess their skill to be above the payoff-relevant cutoff – in this case the median performance – and should vote for the lottery otherwise. Under this assumption, a subject can be classified as accurately calibrated if they: (*i.*) voted for the test and their performance was in the upper half of their group's performance distribution, or – equivalently – (*ii.*) voted for the lottery and their performance was in the lower half of their group's performance distribution[47].

Based on this criterion, 68.4% of participants made an accurate voting decision, representing a significant deviation from a fully accurately calibrated population (One-sided Binomial test: $p = 0.0000$). A breakdown of calibration accuracy across multiple sub-groups of the sample is presented in Table 27.

Table 27: Calibration Accuracy, by Sub-groups

	Fraction	Pearson χ^2-Test
Sub-group		
Vote		
Test	0.641	$\chi^2(1) = 5.63, p = 0.018$
Lottery	0.788	
Gender		
Men	0.704	$\chi^2(1) = 0.65, p = 0.422$
Women	0.658	
Student status		
Students	0.693	$\chi^2(1) = 0.05, p = 0.818$
Non-students	0.679	
Degree in higher education		
Yes	0.731	$\chi^2(1) = 4.37, p = 0.037$
No	0.610	
Sample question performance		
0 correct	0.900	
1 correct	0.590	$\chi^2(3) = 5.78, p = 0.123$
2 correct	0.636	
3 correct	0.726	

Lottery voters were more accurately calibrated than test voters, meaning their vote was more consistent with their subsequent performance. Among those who voted for the test, 64.1% actually performed better than the median, whereas 78.8% of those who voted for the lottery performed worse than the median of their respective group. This statistically significant difference aligns with the

[47] In case a participant's performance was exactly at the median, they were also classified as accurately calibrated.

tendency of participants to overestimate rather than underestimate their own performance in absolute terms in the predictions.

Unlike for voting behavior, the proportion of accurately calibrated participants did not differ significantly by gender or student status. Also, there was no significant difference in age distributions between well-calibrated and inaccurately calibrated participants (mean age of well-calibrated: 33.4, Std. dev. = 12.7; mean age of inaccurately calibrated: 36.3, Std. dev. = 13.2; Two-sample Mann–Whitney test: $|z| = 1.14, p = 0.2541$). However, participants with a university degree were significantly better calibrated – by approximately 12 percentage-points – than those without a degree.

Additionally, calibration accuracy followed a slight U-shaped distribution across different sample performance levels, with participants who performed either very well (all questions correct) or very poorly (none correct) being more accurately calibrated than those with intermediate performance.

As it may be expected, accurately calibrated subjects were significantly more sure about having made the correct choice in the vote after the test than inaccurately calibrated subjects (Two-sample Mann-Whitney U-test: $|z| = 4.22, p < 0.0001$), reporting averages of 5.2 vs. 4.2 on a 7-point scale. Before the test, however, certainty levels did not differ significantly (Two-sample Mann-Whitney U-test: $|z| = 1.42, p = 0.1564$). This suggests that learning their individual test scores provided participants with a clearer sense of whether they had assessed their own abilities correctly, even though they still lacked information on their competitors' performance. Fittingly, inaccurately calibrated subjects reported a significantly higher perceived task difficulty after completing the test compared to accurately well-calibrated subjects (6.0 vs. 5.5 on a 7-point scale;, Two-sample Mann-Whitney U-test: $|z| = 3.85, p = 0.0001$), but not before (Two-sample Mann-Whitney U-test: $|z| = 0.90, p = 0.3706$). They were also significantly less satisfied with their test outcome than well-calibrated subjects (3.3 vs. 4.5, Two-sample Mann-Whitney U-test: $|z| = 4.61, p = 0.0000$).

5.5.4 Self-reported Voting Motives

Immediately after casting their vote in the experiment, subjects were asked to provide a short rationale for their choice. This allows for a deeper understanding of whether individuals followed the equilibrium strategy outlined by H&R or had alternative – monetary or non-monetary – motives underlying their vote that may also explain inaccurate calibration. This section offers an exploratory-descriptive analysis of these self-reported rationales, grouping them into inductively generated thematic categories reflecting latent meanings. [Braun and Clarke, 2006]. The relative frequencies of these motive categories across the full sample are displayed in Figure 19. Example statements for each of the main rationales can be found in Table 28[48]. It is important to note that these self-reported motives are

[48]The complete list of subjects' self-reported rationales and their categorization can be obtained from `https://osf.io/ucx53`.

not necessarily mutually exclusive (e.g., confidence may be altered by signals through sample question outcomes). Therefore, the observed proportions should be interpreted with caution and regarded as indicative tendencies rather than precise measurements. Consequently, statistical tests and cross-comparisons are deliberately avoided.

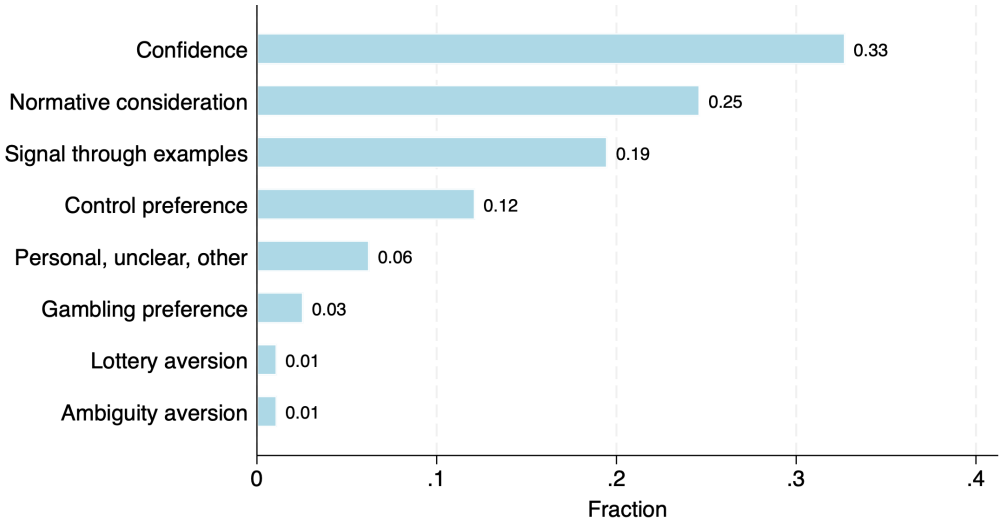


Figure 19: Voting rationales, Self-reported by Subjects (n=272)

According to their responses, approximately one-third of subjects' voting decisions were driven by confidence in their abilities and performance expectations. Differentiating between high and low confidence (see Figure 20) – high confidence resulting in test votes and low confidence in lottery votes – around 35% of test voters and 26% of lottery voters attributed their decision to confidence. A preference for control – in the sense of the potential bonus payment depending on personal performance rather than luck [see Owens et al., 2014; Benoît et al., 2022] – was cited by 17% of test voters (equivalent to 12% of all participants), while – logically – no lottery voters reported this rational. Beyond these two rationales, that were already discussed in the literature (see Appendix D.1), approximately one-fifth of participants reported basing their decision on their sample question performance, with this factor being more prevalent among lottery voters (31%) than test voters (15%). These patterns aligns with the significant differences in voting behavior by sample question outcomes noted in Section 5.5.3, where preference for the test progressively increased with better sample performance. The possibility that sample question results could serve as a decision-making signal was not considered in H&R. Additionally, normative beliefs played a substantial role, with meritocratic beliefs reported by 26% of test voters and preferences for equal chances cited by 22% of lottery voters. As suggested by Krawczyk [2012], notions of fairness were invoked to justify both perspectives [see also Cappelen et al., 2007]: some test voters viewed performance-based allocation as the fairest option, even if they did not expect to receive a bonus themselves. Meanwhile, lottery voters considered it fair if all subjects had

equal statistical chances of receiving the bonus. This motive category appears otherwise unrecognized in the related literature on performance-based versus lottery-based selection mechanisms. The cited lottery rational aligns with this study's arguments made in Section 5.3. Notably, with 36%, normative beliefs represent the most common rational among inaccurately calibrated subjects (see Figure 27 in Appendix D.2), re-emphasizing the role of non-monetary motives leading to – apparently – inaccurate self-assessments in the vote from an equilibrium strategy perspective.

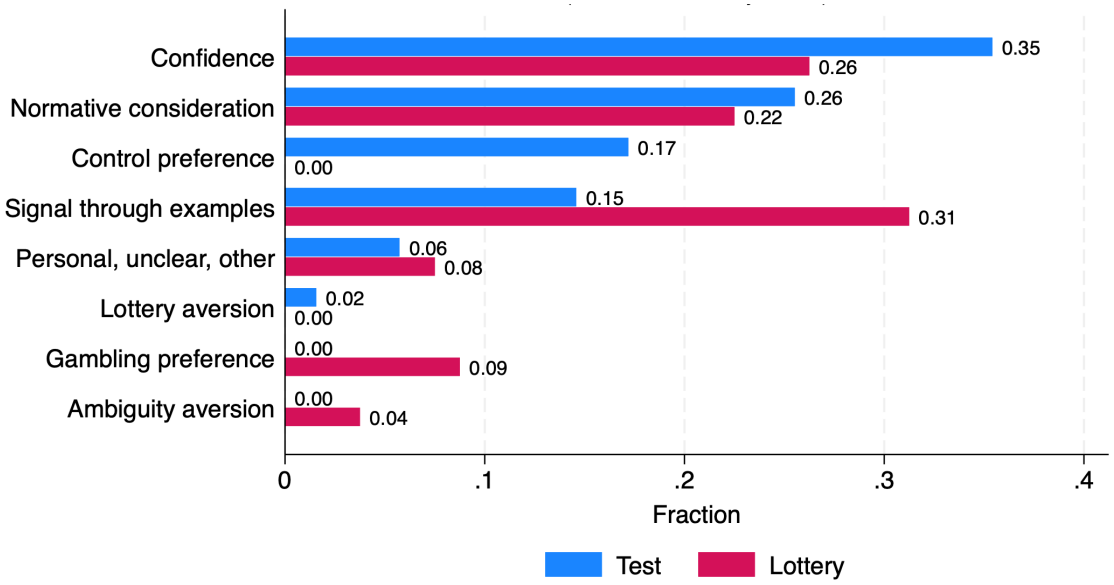


Figure 20: Voting Rationales by Choice in Vote, Self-reported by Subjects (Test: n = 192; Lottery: n = 80)

Other motives, such as ambiguity aversion and preferences for or against gambling, were mentioned only occasionally[49].

While confidence remains the most frequently cited rationale, the data clearly highlight the prominence of alternative non-monetary considerations, such as control preferences and normative beliefs. Moreover, the prevalence of confidence as the voting rationale may still be overstated, not only due to the non-exclusive nature of the response categorization but also because confidence itself is less clearly defined than other motives. For instance, some of the self-reported low confidence among lottery voters may stem from ambiguity aversion without participants explicitly stating it or even being aware of it. Unlike well-articulated beliefs such as fairness considerations or control preferences, confidence represents a more latent sentiment that sometimes serves as a catch-all explanation or a diagnosis of exclusion for various intentional or subconscious behavioral influences [Merkle and Weber, 2011].

[49]Logically, not all rationales apply to both voting outcomes: a preference for control and lottery aversion were only reported by test voters, while ambiguity aversion – a possible motive for choosing the lottery discussed by H&R – and a preference for gambling were exclusive to lottery voters.

Table 28: Example Statements for Self-reported Voting Rationaes, by categorization

Rationale	Example
Confidence	
High	"I believe that I [can/will] deliver an above-average performance and therefore have the best chance of receiving the bonus payment with the performance variant."
Low	"Not enough confidence in my own abilities to definitely finish in the top 50."
Preference for control	
	"I chose 'Test' because I prefer the bonus to depend on my own ability and performance rather than chance/luck."
Sample question signal	
Positive	"I got all the sample tasks correct. If the upcoming tasks are similarly easy/difficult, I have a good chance of finishing in the top 50."
Negative	"I only got 1 out of 3 [sample questions] correct, logically the probabilities are higher if I take the lottery."
Normative beliefs	
Principle of merit	"Those who have better knowledge should be rewarded accordingly. I don't expect to be one of the 50 people to listen, but I think the principle is fair."
Equal chances	"People should have opportunities regardless of their performance. Performance as a factor is only fair if everyone has the same starting conditions. This is never the case in reality."

5.6 Discussion

This study aimed to replicate the findings by H&R and examine their robustness against an alternative lottery mechanism design. By translating their original experiment to an online environment and extending it with standardized questionnaires and a self-reported rationale statement, it may be inferred whether miscalibrated voting distributions arise from inaccurate self-assessments or other factors when individuals choose between a performance-based and a random device-based payoff scheme.

The results reveal voting distributions more consistent with traditional notions of overplacement than with the underplacement pattern found in H&R, as – clearly – no underplacement is observed in any group. The alternative fixed-outcome distribution lottery mechanism is not found to impact voting outcomes. Regardless of treatment, approximately nearly three-fourth of participants preferred the performance-test over the lottery for determining the bonus payment allocation. Correspondingly, participants tended to overestimate their absolute test performance rather than underestimate it.

Generally, the same predictors of voting behavior identified in H&R – predicted own performance, predicted group performance, and the expectation of outperforming the group average ("better") – are also observed here, with predicted own performance being the most influential. Beyond H&R's findings, sample question performance, gender, and student status are additionally found to be significantly related to voting behavior. Notably, sample performance serves as a strong signal for decision-making, with preference for the test over the lottery increasing progressively with the number of correctly

solved sample questions – an effect not explicitly acknowledged in H&R. Meanwhile, social comparison tendencies, general self-efficacy, altruism, risk attitude, ambiguity aversion, and other demographic characteristics show no significant relationship with voting behavior.

An exploratory analysis of participants' self-reported voting rationales reveals four primary drivers of voting behavior. Besides confidence – whether high or low – participants frequently cite their sample question performance as a basis for their decision, reinforcing the correlation patterns observed in this study. While this approach is not strictly rational from an objective perspective – since it neglects the competitive aspect of performance assessment [Hoelzl and Rustichini, 2005; Alicke and Govorun, 2005; Kruger et al., 2008] – it appears to play a substantial role in participants' voting decisions. Additionally, many test voters express a preference for control, indicating a desire to actively influence their payoff basis through the test, despite its inherent ambiguity due to the bonus payment being contingent on others' performance. Finally, a substantial proportion of participants state normative beliefs as their main voting motive. This applies to both choice alternatives, with test voters referencing meritocratic preferences and lottery voters citing equality of opportunity. These self-reported rationales, at least anecdotally, align with the argument put forth in this study, even though its hypotheses are not explicitly supported. Since the latter two rationales – control preferences and normative beliefs – are not grounded in payoff-maximizing considerations, they contradict the equilibrium strategy postulated by H&R.

Overall, participants' behavior and self-assessments appear reasonably consistent, with the calibration of above- and below-median performers opting for the test and lottery, respectively, being relatively accurate at close to 70%. While using sample performance as a proxy for test performance is not objectively rational – since it conflates absolute and relative evaluations – it is a subjectively understandable heuristic, as it serves as the only available reference point for participants. While the present sample is more demographically diverse than H&R's original sample, this does not seem to have significant implications – except, perhaps, that students show a greater inclination toward voting for the test than non-students. However, this trend is not clearly reflected in comparisons with H&R's results, as their sample includes a larger proportion of students.

Limitations of this study concern the comparability between the performance test used in the original study and the alternative task employed in this replication. Using the task from H&R's 2005 experiment – a vocabulary-based knowledge test – in its original form would have been impractical under modern technological conditions, both in online and lab settings, due to the prevalence of mobile devices giving participants the opportunity of searching for answers online. Even with modifications such as a timer, the informative value of the results would have been limited. Consequently, this study replaces it with an alternative format – logical analogy selection under time pressure – designed to en-

gage similar cognitive abilities. Since anticipated task difficulty, which is difficult to assess in isolation, plays a crucial role in expected self-assessment miscalibration, any divergence between this study's results and those of the original should be interpreted with caution. Based on average performance, this task's difficulty falls between H&R's "difficult" and "easy" tasks. Thus, it may be assumed that participants did not perceive it as genuinely "difficult" (H&R's *Difficult* groups solved an average of 7 items). However, participants in this study rated the task as more difficult than subjects in H&R even before taking the test, and significantly adjusted their stated difficulty perceptions upward post-test – despite performing well on both sample questions and the test itself. Some explicitly emphasized task difficulty in their self-reported rationales.

A key issue is the absence of an objective measure of difficulty, as H&R differentiate task difficulty only in relative terms. Moreover, they report only the average perceived task difficulty across all participants, despite having a design that explicitly distinguishes between two difficulty levels. In the current study, a logit regression of voting behavior on pre-test difficulty assessments yields a non-significant coefficient that is barely different from zero ($\beta 1 = -0.102, |z| = 0.63, p = 0.531$; M.E.: $-.0219, p = 0.529$), suggesting that factors other than task difficulty play a more central role in voting decisions – consistent with the study's findings. H&R report a similar non-effect in all groups except the *Difficult* $\times$ *Money* condition, where they find a counterintuitive positive and significant coefficient, implying an increasingly strong preference for the test the more difficult it is perceived under monetary incentives. However, H&R do not further discuss potential reasons for this effect. Furthermore, the task in this study offered a ten-fold higher chance of success through guessing. However, in principle, this should not affect relative overconfidence. To counterbalance this and prevent participants from searching for answers online, time pressure was introduced to increase the perceived difficulty of the task.

Apart from the different task, two other minor factors distinguish this replication from the original experiment (see also Table 52 in Appendix D.2), primarily for operational reasons. First, the increased physical distance between participants in an online environment may have influenced decision-making. Better-than-average effects are usually found to be stronger when people compare themselves to an abstract group rather than to a specific known individual [Moore and Cain, 2007], and tend to weaken with personal contact or when comparison targets become more individuated and concrete [Alicke et al., 1995]. This distinction may partially apply to the contrast between centralized lab settings and remote participation, potentially leading to differences in voting behavior — even though no specific information about comparison targets was provided in either case, and direct interaction between subjects was absent in the lab as well. However, online experiments have generally been found to provide "adequate and reliable" [Arechar et al., 2017, p. 100] alternatives to lab-based studies, with

both approaches yielding consistent and robust results [Prissé and Jorrat, 2022]. Second, this study featured a significantly larger reference group – 100 participants in each group – compared to the original experiment, where subjects were placed in smaller groups of "at least nine". While a larger reference group should not impact a subject's relative self-assessment from an equilibrium perspective, empirical findings suggest that an increased number of competitors can reduce competitive motivation [Garcia and Tor, 2009]. However, this should have promoted lottery votes rather than test votes, which was not the case in this experiment. Consequently, the reasons for the observed differences in voting behavior between H&R and this study remain largely unclear at this stage.

As in H&R, the task in this study primarily assessed linguistic competencies, combined with elements of logical reasoning. Future research could further reinforce these findings by incorporating alternative test formats, such as mathematical problems or matrix-based tasks, as these may elicit different miscalibration dynamics across academic domains [Erickson and Heit, 2015]. Furthermore, this study, like H&R, does not cover motivational and status-based [see e.g., Alicke and Govorun, 2005; Anderson et al., 2012; Brown, 2012; Bénabou and Tirole, 2002] or social [see e.g., Proeger and Meub, 2014; Tenney et al., 2019; Taylor and Brown, 1988] accounts of overconfidence. These perspectives are effectively abstracted from the analysis and could provide valuable avenues for future research.

Replicating the classic experiment by H&R, this study contributes to ongoing methodological discussions in experimental overconfidence research on the validity of using choices between performance-based and lottery-based payoffs as behavioral measures of overconfidence, as well as the broader conceptual validity of the overconfidence paradigm [see e.g., Benoît and Dubra, 2011; Merkle and Weber, 2011; Owens et al., 2014]. Specifically, it seeks to gain a more comprehensive understanding of the factors contributing to misplacement in performance self-assessments beyond overconfidence and to explore alternative explanations for apparent overconfidence in overplacement data, thereby clarifying the scope of phenomena encompassed within the term "overconfidence".

Bibliography

A., S. and R., S. (2023). A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. Decision Analytics Journal, 7:100230.

Abbink, K., Irlenbusch, B., and Renner, E. (2002). An experimental bribery game. The Journal of Law, Economics and Organization, 18(2):428–454.

Abeler, J., Altmann, S., Kube, S., and Wibral, M. (2010). Gift exchange and workers' fairness concerns: When equality is unfair. Journal of the European Economic Association, 8(6):1299–1324.

Abeler, J., Becker, A., and Falk, A. (2014). Representative evidence on lying costs. Journal of Public Economics, 113:96–104.

Abeler, J., Nosenzo, D., and Raymond, C. (2019). Preferences for truth-telling. Econometrica, 87(4):1115–1153.

Akerlof, G. A. and Kranton, R. E. (2000). Economics and identity. Quarterly Journal of Economics, 115:715–753.

Albaba, B. M. and Yildiz, Y. (2019). Modeling cyber-physical human systems via an interplay between reinforcement learning and game theory. Annual Reviews in Control, 48(1–19).

Albert, D. A. and Smilek, D. (2023). Comparing attentional disengagement between Prolific and MTurk samples. Scientific Reports, 13(1):20574.

Alicke, M. D. and Govorun, O. (2005). The better-than-average effect. In The Self in Social Judgement, pages 85–106. Psychology Press.

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J., and Vredenburg, D. S. (1995). Personal contact, individuation, and the better-than-average effect. Journal of Personality and Social Psychology, 68(5):804–825.

Andersen, S., Harrison, G. W., Lau, M. I., and Rutström, E. E. (2006). Elicitation using multiple price list formats. Experimental Economics, 9:383–405.

Anderson, C., Brion, S., Moore, D. A., and Kennedy, J. A. (2012). A status-enhancement account of overconfidence. Journal of Personality and Social Psychology, 103(4):718–735.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. The Economic Journal, 100(401):464–477.

Andreoni, J. and Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. Econometrica, 70(2):737–753.

Arechar, A. A., Gächter, S., and Molleman, L. (2017). Conducting interactive experiments online. Experimental Economics, 21:99–131.

Ariely, D., Kamenica, E., and Prelec, D. (2008). Man's search for meaning: The case of legos. Journal of Economic Behavior & Organization, 67(3-4):671–677.

Arkes, H. R., Dawes, R. M., and Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. Organizational Behavior and Human Decision Processes, 37:93–110.

Aron, A., Aron, E. N., and Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. Journal of Personality and Social Psychology, 63(4):596–612.

Ayton, P. and Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? Memory & Cognition, 32(8):1369–1378.

Babina, T., Fedyk, A., He, A., and Hodson, J. (2024). Artificial intelligence, firm growth, and product innovation. Journal of Financial Economics, 151:103745.

Baek, H. Y. and Neymotin, F. (2019). Overconfident entrepreneurs: Innovating more and paying the piper. Economics Bulletin, 39:1144–1153.

Baghdasaryan, V., Davtyan, H., Sarikyan, A., and Navasardyan, Z. (2022). Improving tax audit efficiency using machine learning: The role of taxpayer's network data in fraud detection. Applied Artificial Intelligence, 36(1):e2012002.

Bakumenko, A. and Elragal, A. (2022). Detecting anomalies in financial data using machine learning algorithms. Systems, 10:130.

Banker, S. and Khetani, S. (2019). Algorithm overdependence: How the use of algorithmic recommendation systems can increase risks to consumer well-being. Journal of Public Policy & Marketing, 38:500–515.

Bar-Shalom, Y. and Tse, E. (1974). Dual effect, certainty equivalence, and separation in stochastic control. IEEE Transactions on Automatic Control, 19(5):494– 500.

Barber, B. M. and Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. The Journal of Finance, 55(2):773–806.

Barber, B. M. and Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. The Quarterly Journal of Economics, 116(1):261–292.

Bartling, B. and Fischbacher, U. (2012). Shifting the blame: On delegation and responsibility. Review of Economic Studies, 79(1):67–87.

Batson, C. D. and Shaw, L. L. (1991). Evidence for altruism: Toward a pluralism of prosocial motives. Psychological Inquiry, 2(2):107–122.

Beierlein, C., Kemper, C. J., Kovaleva, A., and Rammstedt, B. (2013). Short scale for measuring general self-efficacy beliefs (ASKU). methods, data, analyses, 7(2):251–278.

Ben-David, I., Graham, J. R., and Harvey, C. R. (2013). Managerial miscalibration. The Quarterly Journal of Economics, 128(4):1547–1584.

Bénabou, R. and Tirole, J. (2002). Self-confidence and personal motivation. The Quarterly Journal of Economics, 117(3):871–915.

Bénabou, R. and Tirole, J. (2011). Identity, morals, and taboos: Beliefs as assets. The Quarterly Journal of Economics, 126(2):805–855.

Benartzi, S. (2001). Excessive extrapolation and the allocation of 401(k) accounts to company stock. The Journal of Finance, 56(5):1747–1764.

Benoît, J.-P. and Dubra, J. (2011). Apparent overconfidence. Econometrica, 79(5):1591–1625.

Benoît, J.-P., Dubra, J., and Moore, D. (2014). Does the better-than-average effect show that people are overconfident?: Two experiments. Journal of the European Economic Association, 13(2):293–329.

Benoît, J.-P., Dubra, J., and Romagnoli, G. (2022). Belief elicitation when more than money matters: Controlling for 'control'. American Economic Journal: Microeconomics, 14(3):837–888.

Bernardo, A. E. and Welch, I. (2001). On the evolution of overconfidence and entrepreneurs. Journal of Economics & Management Strategy, 10(3):301–330.

Bertsekas, D. P. (2005). Dynamic Programming and Optimal Control, Vol.1. Athena Scientific, Belmont, MA.

Bertsekas, D. P. (2019). Reinforcement Learning and Optimal Control. Athena Scientific, Belmont, MA.

Biais, B., Hilton, D., Mazurier, K., and Pouget, S. (2005). Judgemental overconfidence, self-monitoring, and trading performance in an experimental financial market. Review of Economic Studies, 72(2):287–312.

Bicchieri, C. (2006). Norm nudging and twisting preferences. Behavioural Public Policy, 7:914–923.

Bicchieri, C., Dimant, E., and Sonderegger, S. (2023). It's not a lie if you believe the norm does not apply: Conditional norm-following and belief distortion. Games and Economic Behavior, 138:321–354.

Bicchieri, C. and Xiao, E. (2009). Do the right thing: But only if others do so. Journal of Behavioral Decision Making, 22(2):191–208.

Bickley, S. J., Chan, H. F., and Torgler, B. (2022). Artificial intelligence in the field of economics. Scientometrics, 127:2055–2084.

Biener, C. and Waeber, A. (2024). Would I lie to you? How interaction with chatbots induces dishonesty. Journal of Behavioral and Experimental Economics, 112:102279.

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

Bignami, F. (2022). Artificial intelligence accountability of public administration. The American Journal of Comparative Law, 70(Issue Supplement 1):i312–i346.

Billett, M. T. and Qian, Y. (2008). Are overconfident ceos born or made? Evidence of self-attribution bias from frequent acquirers. Management Science, 54(6):1037–1051.

Bingley, W. J., Curtis, C., Lockey, S., Bialkowski, A., Gillespie, N., Haslam, S. A., Ko, R. K., Steffens, N., Wiles, J., and Worthy, P. (2023). Where is the human in human-centered AI? insights from developer priorities and user experiences. Computers in Human Behavior, 141:107617.

Black, E., Elzayn, H., Chouldechova, A., Goldin, J., and Ho, D. E. (2022). Algorithmic fairness and vertical equity: Income fairness with IRS tax audit models. ACM Conference on Fairness, Accountability and Transparency,June 21–24, 2022, Seoul, Republic of Korea, pages 1479–1503.

Blais, A.-R. and Weber, E. U. (2006). A domain-specific risk-taking (DOSPERT) scale for adult populations. Management Science, 1(1):33–47.

Blavatskyy, P. and de Roos, N. (2023). Large-stakes estimates of risk and ambiguity attitudes. SSRN Discussion Paper No. 3740218.

Blavatskyy, P. R. (2009). Betting on own knowledge: Experimental test of overconfidence. Journal of Riska and Uncertainty, 38:39–49.

Bogert, E., Schecter, A., and Watson, R. T. (2021). Humans rely more on algorithms than social influence as a task becomes more difficult. Scientific Reports, 11(8028).

Bolle, F. (1990). High reward experiments without high expenditure for the experimenter. Journal of Economic Psychology, 11(2):157–167.

Bolton, G., Dimant, E., and Schmidt, U. (2021). Observability and social image: On the robustness and fragility of reciprocity. Journal of Economic Behavior & Organization, 191:946–964.

Bolton, G. E. and Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. The American Economic Review, 90(1):166–193.

Bolton, G. E. and Ockenfels, A. (2012). Behavioral economic engineering. Journal of Economic Psychology, 33(3):665–676.

Borges, B., Goldstein, D. G., Ortmann, A., and Gigerenzer, G. (1999). Can ignorance beat the stock market. In Simple heuristics that make us smart. Oxford University Press.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3:77–101.

Brown, J. D. (2012). Understanding the better than average effect: Motives (still) matter. Journal of Personality and Social Psychology, 38(2):209–219.

Brown, J. L., Farrington, S., and Sprinkle, G. B. (2016). Biased self-assessments, feedback, and employees' compensation plan choices. Accounting, Organizations and Society, 54:45–59.

Brown, W. O. and Sauer, R. D. (1993). Does the Basketball Market Believe in the Hot Hand? Comment. The American Economic Review, 83(5):1377–1386.

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? a tutorial of power analysis with reference tables. Journal of Cognition, 2(1):1–38.

Burkart, N. and Huber, M. F. (2021). A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research, 70:245–317.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society, 3:1–12.

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

Butters, J. (2025). More than 40Technical report, FactSet Insight.

Cain, D. M., Moore, D. A., and Haran, U. (2015). Making sense of overconfidence in market entry. Strategic Management Journal, 36(1):1597–1597.

Camerer, C. F. (1989). Does the basketball market believe in the 'hot hand'? The American Economic Review, 79(5):1257–1261.

Camerer, C. F. (1999). Behavioral economics: Reunifying psychology and economics. Proceedings of the National Academy of Sciences of the United States of America, 96:10575–10577.

Camerer, C. F. (2019). Artificial intelligence and behavioral economics. In The Economics of Artificial Intelligence: An Agenda. University of Chicago Press.

Camerer, C. F., Ho, T.-H., and Chong, J.-K. (2004). A cognitive hierarchy model of games. The Quarterly Journal of Economics, 119(3).

Camerer, C. F. and Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. Journal of Risk and Uncertainty, 19(1-3):7–42.

Camerer, C. F. and Loewenstein, C. (2003). Behavioral economics: Past, present, future. In Camerer, C. F., Loewenstein, G., and Rabin, M., editors, Advances in Behavioral Economics, pages 3–52. Princeton, NJ: Princeton University Press.

Camerer, C. F. and Loewenstein, G. (2004). Behavioral economics: Past, present and future. In Advances in Behavioral Economics. Russel Sage Foundation.

Camerer, C. F. and Lovallo, D. (1999). Overconfidence and excess entry: An experimental approach. The American Economic Review, 89(1):306–318.

Canning, C., Donahue, T. J., and Scheutz, M. (2014). Investigating human perceptions of robot capabilities in remote human-robot team tasks based on first-person robot video feeds. International Conference on Intelligent Robots and Systems (IROS 2014), pages 4354–4361.

Cappelen, A. W., Sørensen, E. Ø., and Tungodden, B. (2007). The pluralism of fairness ideals: An experimental approach. American Economic Review, 97(3):818–827.

Carbonara, E., Santarelli, E., and Tripathi, I. (2025). Assessing the impact of AI on labor market outcomes: A meta-analysis. SSRN Discussion paper no. 5126345. https://ssrn.com/abstract=5126345.

Carlson, K. A. and Shu, S. B. (2007). The rule of three: How the third event signals the emergence of a streak. Organizational Behavior and Human Decision Processes, 104(1):113—121.

Carrington, M. J., Neville, B. A., and Whitwell, G. J. (2010). Why Ethical Consumers Don't Walk Their Talk: Towards a Framework for Understanding the Gap Between the Ethical Purchase Intentions and Actual Buying Behaviour of Ethically Minded Consumers. Journal of Business Ethics, 97(1):139–158.

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

Chander, A., Srinivasan, R., Chelian, S., Wang, J., and Uchino, K. (2018). Working with beliefs: AI transparency in the enterprise. IUI Workshops.

Charness, G. (2010). Laboratory experiments: Challenges and promise. a review of "theory and experiment: What are the questions?" by Vernon Smith. Journal of Economic Behavior & Organization, 73:21–23.

Charness, G., Gneezy, U., and Halladay, B. (2016). Experimental methods: Pay one or pay all. Journal of Economic Behavior & Organization, 131(A):141–150.

Charness, G. and Levin, D. (2005). When optimal choices feel wrong: A laboratory study of bayesian updating, complexity and affect. The American Economic Review, 95(4):1300–1309.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. The Quarterly Journal of Economics, 117(3):817–869.

Chau, A. W. and Phillips, J. G. (1995). Effects of perceived control upon wagering and attributions in computer blackjack. The Journal of General Psychology, 122(3):253–269.

Chen, D. L., Schonger, M., and Wickens, C. (2016). oTree—an open-source platform for laboratory, online, and field experiments. Journal of Behavioral and Experimental Finance, 9:88–97.

Chen, G., Crossland, C., and Luo, S. (2015). Making the same mistake all over again: Ceo overconfidence and corporate resistance to corrective feedback. Strategic Management Journal, 36(10):1513–1535.

Cheng, J.-Z., Ni, D., Chou, Y.-H., Qin, J., Tiu, C.-M., Chang, Y.-C., Huang, C.-S., Shen, D., and Chen, C.-M. (2016). Computer-aided diagnosis with deep learning architecture: Applications to breast lesions in us images and pulmonary nodules in ct scans. Scientific Reports, 6:24454.

Chow, C. C. and Sarin, R. K. (2001). Comparative ignorance and the ellsberg paradox. Journal of Risk and Uncertainty, 22(2):129–139.

Christou, P. A. (2023). How to use artificial intelligence (AI) as a resource, methodological and analysis tool in qualitative research? The Qualitative Report, 28:1968–1980.

Chuang, W.-I. and Lee, B.-S. (2006). An empirical evaluation of the overconfidence hypothesis. Journal of Banking & Finance, 30(9):2489–2515.

Chui, M., Roberts, R., Yee, L., Hazan, E., Singla, A., Smaje, K., Sukharevsky, A., and Zemmel, R. (2023). The economic potential of generative AI: The next productivity frontier. Technical report, McKinsey & Company.

Church, K. W. (2017). Word2vec. Natural Language Engineering, 23:155–162.

Clark, J. and Friesen, L. (2009). Overconfidence in forecasts of own performance: An experimental study. The Economic Journal, 119(534):229–251.

Coeckelbergh, M. (2020). Artifcial intelligence, responsibility attribution, and a relational justifcation of explainability. Science and Engineering Ethics, 26:2051–2068.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. Academic Press.

Cohen, J. (2013). Statistical power analysis for the behavioral sciences. Routledge.

Cohn, A., Gesche, T., and Maréchal, M. A. (2022). Honesty in the digital age. Management Science, 68(2):809–1589.

Congdon, W. J. and Shankar, M. (2015). The white house social & behavioral sciences team: Lessons learned from year one. Behavioral Science & Policy, 1:77–86.

Croson, R. (2005). The method of experimental economics. International Negotiation, 10:131–148.

Croson, R. and Sundali, J. (2005). The gambler's fallacy and the hot hand: Empirical data from casinos. The Journal of Risk and Uncertainty, 30(3):195–209.

Daniel, K. and Hirshleifer, D. (2015). Overconfident and investors, predictable returns, and excessive trading. Journal of Economic Perspectives, 29(4):61–88.

Daniel, K. D., Hirshleifer, D., and Subrahmanyam, A. (2001). Overconfidence, arbitrage, and equilibrium asset pricing. The Journal of Finance, 56(3):921–965.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. American Psychologist, 34(7):571–582.

Dawes, R. M., Faust, D., and Meehl, P. E. (1989). Clinical versus actuarial judgment. Science, 243(4899):1668–1674.

DeBondt, W. and Thaler, R. H. (1995). Financial decision-making in markets and firms: a behavioral perspective. In Handbooks in Operations Research and Management Science, volume 9, pages 385–410. North Holland.

Denrell, J. and March, J. G. (2001). Adaptation as information restriction: The hot stove effect. Organization Science, 12(5):523–659.

Dietvorst, B. J. and 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., and 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., and Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. Management Science, 64(3):1155–1170.

Djawadi, B. M. and Fahr, R. (2013). The impact of risk perception and risk attitudes on corrupt behavior: Evidence from a petty corruption experiment. IZA Discussion Paper No. 7383.

Djawadi, B. M. and Fahr, R. (2015). "... and they are really lying": Clean evidence on the pervasiveness of cheating in professional contexts from a field experiment. Journal of Economic Psychology, 48:48–59.

Dohmen, T., Falk, A., Huffman, D., and Sunde, U. (2010). Are risk aversion and impatience related to cognitive ability. American Economic Review, 100(3):1238–1260.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., and Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. Journal of the European Economic Association, 9:522–550.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Dressler, F. (2018). Cyber physical social systems: Towards deeply integrated hybridized systems. In Intern. Conf. Computing, Networking and Communications.

Duarte, F. (2025). Number of ChatGPT Users (June 2025). Technical report, Exploding Topics.

Duffy, J. and Ochs, J. (2009). Cooperative behavior and the frequency of social interaction. Games and Economic Behavior, 66(2):785–812.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., and Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. Human Factors, 44(1):79–94.

Dzindolet, M. T., Scott A. Peterson, R. A. P., Pierce, L. G., and Beck, H. P. (2003). The role of trust in automation reliance. Econometrica, 58(4):697–718.

Eckel, C. C. and Grossman, P. J. (1996). Altruism in anonymous dictator games. Games and Economic Behavior, 16:181–191.

Eckel, C. C. and Grossman, P. J. (2008). Men, women and risk aversion: Experimental evidence. In Plott, C. R. and Smith, V. L., editors, Handbook of Experimental Economic Results, volume 1, chapter 113, pages 1061–1073. Amsterdam: Elsevier B.V.

Einhorn, H. J. and Hogarth, R. M. (1986). Decision making under ambiguity. The Journal of Business, 4(2):S225–S250.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. The Quarterly Journal of Economics, 75(4):643–669.

Engelmann, D. and Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. The American Economic Review, 94(4):857–869.

Ercan, Z., Carvalho, A., Gokasan, M., and Borelli, F. (2017). Modeling, identification, and predictive control of a driver steering assistance system. IEEE Trans. Human-Machine Systems, 47(5):700–710.

Erickson, S. and Heit, E. (2015). Metacognition and confidence: comparing math to other academic subjects. Frontiers in Psychology, 6:742.

Ericson, K. M. M. (2011). Forgetting we forget: Overconfidende and memory. Journal of the European Economic Association, 9(1):43–60.

European Commission (2019). Launch of the competence centre on behavioural insights. https://knowledge4policy.ec.europa.eu/event/launch-competence-centre-behavioural-insights [accessed 07.07.2025].

Eyssel, F. and Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. British Journal of Social Psychology, 51:724–731.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and Sunde, U. (2018). Global evidence on economic preferences. The Quarterly Journal of Economics, 133:1645–1692.

Falk, A. and Fischbacher, U. (2006). A theory of reciprocity. Games and Economic Behavior, 54:293–315.

Falk, A. and Heckman, J. J. (2022). Lab experiments are a major source of knowledge in the social sciences. Science, 326(5952):535–538.

Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41(4):1149–1160.

Faúndez-Ugalde, A., Mellado-Silva, R., and Aldunate-Lizana, E. (2020). Use of artificial intelligence by tax administrations: An analysis regarding taxpayers' rights in latin american countries. Computer Law & Security Review, 38:105441.

Fehr, E. and Fischbacher, U. (2004). Social norms and human cooperation. Trends in Cognitive Sciences, 8:185–190.

Fehr, E. and Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. Journal of Economic Perspectives, 14:159–181.

Fehr, E., Naef, M., and Schmidt, K. M. (2006). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments: Comment. The American Economic Review, 96(5):1912–1917.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. The Quarterly Journal of Economics, 114(3):817–868.

Feng, L., Wiltsche, C., Humphrey, L., and Topcu, U. (2016). Synthesis of human-in-the-loop control protocols for autonomous systems. IEEE Transactions on Automation Science and Engineering, 13(2):450–462.

Fenneman, A., Sickmann, J., Pitz, T., and Sanfey, A. G. (2021). Two distinct and separable processes underlie individual differences in algorithm adherence: Differences in predictions and differences in trust thresholds. PLOS ONE, 16(2).

Ferguson, T. S. (2008). Optimal stopping and applications. Mathematics Department, UCLA, https://www.math.ucla.edu/people/ladder/tom.

Filiz, I., Judek, J., Lorenz, M., and Spiwoks, M. (2021). The tragedy of algorithm aversion. Wolfsburg Working Papers No. 21-02.

Fischbacher, U. and Föllmi-Heusi, F. (2013). Lies in disguise - An experimental study on cheating. Journal of the European Economic Association, 11(3):525–547.

Fischbacher, U., Kairies-Schwarz, N., and Stefani, U. (2017). Non-additivity and the salience of marginal productivities: Experimental evidence on distributive fairness. The Economic Journal, 127(603):587–610.

Fox, C. R. and Tversky, A. (1995). Ambiguity aversion and comparative ignorance. The Quarterly Journal of Economics, 110(3):585–603.

Franke, T., Attig, C., and Wessel, D. (2018). A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ATI) scale. International Journal of Human–Computer Interaction, 35(6):456–467.

Freyer, T. and Günther, L. R. K. (2023). Inherited inequality and the dilemma of meritocracy. ECONtribute Discussion Paper No. 171.

Friedman, D. and Sunder, S. (1994). Experimental methods: A primer for economists. Cambridge University Press.

Frieze, I. and Weiner, B. (1971). Cue utilization and attributional judgments for success and failure. Journal of Personality, 39(4):591–605.

Fuchs, C., Matt, C., Hess, T., and Hoerndlein, C. (2016). Human vs. algorithmic recommendations in big data and the role of ambiguity. AMCIS 2016: Surfing the IT Innovation Wave - 22nd Americas Conference on Information Systems.

Garcia, S. M. and Tor, A. (2009). The N-effect: More competitors, less competition. Strategic Management Journal, 20(7):871–877.

Gee, L. K., Migueis, M., and Parsa, S. (2017). Redistributive choices and increasing income inequality: Experimental evidence for income as a signal of deservingness. Experimental Economics, 20(4):894–923.

Gerlach, P., Teodorescu, K., and Hertwig, R. (2019). The truth about lies: A meta-analysis on dishonest behavior. Psychological Bulletin, 145(1):1–44.

Gervais, S. and Odean, T. (2001). Learning to be overconfident. The Review of Financial Studies, 14(1):1–27.

Ghosh, D. and Ray, M. R. (1997). Risk, ambiguity, and decision choice: Some additional evidence. Decision Sciences, 28(1):81–104.

Gibbons, F. X. and Buunk, B. P. (1999). Individual differences in social comparison: Development of a scale of social comparison orientation. Journal of Personality and Social Psychology, 76(1):129–142.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond 'heurisics and biases'. In European Review of Social Psychology, volume 2. John Wiley & Sons Ltd.

Gigerenzer, G. (1996). On Narrow Norms and Vague Heuristics: A Reply to Kahneman and Tversky. Psychological Review, 103:592–596.

Gigerenzer, G. (2001). The adaptive toolbox. In Bounded Rationality: The Adaptive Toolbox. The MIT Press.

Gigerenzer, G. (2016). Towards a rational theory of heuristics. In Archival Insights into the Evolution of Economics. Palgrave Macmillan.

Gigerenzer, G. and Hoffrage, U. (1995). How to improve bayesian reasoning without instruction: Frequency formats. Psychological Review, 102:684–704.

Gigerenzer, G. and Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In Simple heuristics that make us smart. Oxford University Press.

Gillespie, T. (2016). Algorithm. In Digital Keywords: A Vocabulary of Information Society and Culture. Princeton University Press.

Gilovich, T., Vallone, R., and Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. Cognitive Psychology, 17(3):295–314.

Gino, F., Brooks, A. W., and Schweitzer, M. E. (2012). Anxiety, advice, and the ability to discern: feeling anxious motivates individuals to seek and use advice. Journal of Personality and Social Psychology, 102(3):497–512.

Gneezy, U., Imas, A., and List, J. (2015). Estimating individual ambiguity aversion: A simple approach. National Bureau of Economic Research Working Paper 20982.

Gneezy, U., Kajackaite, A., and Sobel, J. (2018). Lying aversion and the size of the lie. The American Economic Review, 108(2):419–453.

Goel, A. M. and Thakor, A. V. (2008). Overconfidence, CEO Selection, and Corporate Governance. The Journal of Finance, 63(6):2737–2784.

Gogoll, J. and Uhl, M. (2018). Rage against the machine: Automation in the moral domain. Journal of Behavioral and Experimental Economics, 74:97–103.

Goldstein, D. G. and Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. Psychological Review, 109:75–90.

Goodie, A. S. and Young, D. L. (2007). The skill element in decision making under uncertainty: Control or competence? Judgment and Decision Making, 2(3):189–203.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. Journal of the Economic Science Association, 1:114–125.

Grieco, D. and Hogarth, R. M. (2009). Overconfidence in absolute and relative performance: The regression hypothesis and bayesian updating. Journal of Economic Psychology, 30:756–771.

Grossman, Z. and Owens, D. (2012). An unlucky feeling: Overconfidence and noisy feedback. Journal of Economic Behavior & Organization, 84(2):510–524.

Gruber, K. (2019). Is the future of medical diagnosis in computer algorithms? The Lancet Digital Health, 1(1):E15–E16.

Guala, F. (2005). The Methodology of Experimental Economics. Cambridge University Press.

Gubaydullina, Z., Judek, J. R., Lorenz, M., and Spiwoks, M. (2022). Comparing different kinds of influence on an algorithm in its forecasting process and their impact on algorithm aversion. Businesses, 2:448–470.

Güth, W., Schmittberger, R., and Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. Journal of Economic Behavior & Organization, 3:367–388.

Hajkowicz, S., Sanderson, C., Karimi, S., Bratanova, A., and Naughtin, C. (2023). Artificial intelligence adoption in the physical sciences, natural sciences, life sciences, social sciences and the arts and humanities: A bibliometric analysis of research publications from 1960-2021. Technology in Society, 74:102260.

Hao, L. and Houser, D. (2008). Perceptions, intentions, and cheating. Journal of Economic Behavior & Organization, 133:52–73.

Harrison, G. W., Lau, M. I., and Rutström, E. E. (2007a). Estimating risk attitudes in denmark: A field experiment. The Scandinavian Journal of Economics, 109(2):341–368.

Harrison, G. W., List, J. A., and Towe, C. (2007b). Naturally occurring preferences and exogenous laboratory experiments: A case study of risk aversion. Econometrica, 75(2):433–458.

Harrison, G. W., Martínez-Correa, J., and Swarthout, J. T. (2015). Reduction of compound lotteries with objective probabilities: Theory and evidence. Journal of Economic Behavior & Organization, 119:32–55.

Haslam, N. (2006). Dehumanization: An integrative review. Personality and Social Psychology Review, 10(3):252–264.

He, J. and King, W. R. (2008). The role of user participation in information systems development: Implications from a meta-analysis. Journal of Management Information Systems, 25(1):301–331.

He, X., Zhao, K., and Chu, X. (2021). Automl: A survey of the state-of-the-art. Knowledge-Based Systems, 212:106622.

Healy, P. J. and Moore, D. A. (2007). Bayesian overconfidence. SSRN Working Paper No. 1001820.

Heath, C. and Tversky, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. Journal of Risk and Uncertainty, 4:5–28.

Hertwig, R. and Gigerenzer, G. (1999). The 'conjunction fallacy' revisited: How intelligent inferences look like reasoning errors. Journal of Behavioral Decision Making, pages 275–305.

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. Industrial and Organizational Psychology, 1(3):333–342.

Hilary, G. and Menzly, L. (2006). Does past success lead analysts to become overconfident? Management Science, 52(4):489–500.

Hirschman, D. (2016). Stylized facts in the social sciences. Sociological Science, 3:207–232.

Hirshleifer, D., Low, A., and Teoh, S. H. (2012). Are Overconfident CEOs Better Innovators? The Journal of Finance, 67(4):1457–1498.

Hoelzl, E. and Rustichini, A. (2005). Overconfident: Do you put your money on it? The Economic Journal, 115(503):305–318.

Hoffman, R. R., Johnson, M., and Bradshaw, J. M. (2013). Trust in automation. Human-centered Computing, 28(1):84–88.

Hoffrage, U., Lindsey, S., Hertwig, R., and Gigerenzer, G. (2000). Communicating statistical information. Science, 290:2261–2262.

Hogarth, R. M. (1981). Beyond discrete biases: Functional and dysfunctional aspects of judgmental heuristics. Psychological Bulletin, 90(2):197–217.

Hogarth, R. M. and Kunreuther, H. (1989). Risk, ambiguity, and insurance. Journal of Risk and Uncertainty, 2:5–35.

Hokayem, P. F. and Spong, M. W. (2006). Bilateral teleoperation: An historical survey. Automatica, 42(12).

Hollard, G., Massoni, S., and Vergnaud, J.-C. (2016). In search of good probability assessors: an experimental comparison of elicitation rules for confidence judgments. Journal of the European Economic Association, 80:363–387.

Holt, C. A. (1986). Preference reversals and the independence axiom. The American Economic Review, 76:508–515.

Holt, C. A. and Laury, S. K. (2002). Risk aversion and incentive effects. American Economic Review, 92(5):1644–1655.

Holzinger, A., Biemann, C., Pattichis, C. S., and Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? https://doi.org/10.48550/arXiv.1712.09923.

Hou, Y. T.-Y. and Jung, M. F. (2021). Who is the Expert? Reconciling Algorithm Aversion and Algorithm Appreciation in AI-Supported Decision Making. Proceedings of the ACM on Human-Computer Interaction, 5:1–25.

Howard, R. A. and Matheson, J. E. (1972). Risk-sensitive markov decision processes. Management Science, 18(7):365–369.

Inoue, M. and Gupta, V. (2019). Weak control for human-in-the-loop systems. IEEE Control Syst. Lett., 3(2):440–445.

Irlenbusch, B. and Köbis, N. (2025). Behavioural ethics of artificial intelligence. In Elgar Encyclopedia of Behavioural and Experimental Economics. Edward Elgar Publishing Limited.

Jacobsen, C., Fosgaard, T. R., and Pascual-Ezama, D. (2018). Why do we lie? A practical guide to the dishonesty literature. Journal of Economic Surveys, 32(2):357–387.

Jacobsen, C. and Piovesan, M. (2016). Tax me if you can: An artifactual field experiment on dishonesty. Journal of Economic Behavior & Organization, 124:7–14.

Jago, A. S. (2023). Algorithms and authenticity. Academy of Management Discoveries, 5(1):38–56.

Jaquette, S. C. (1976). A utility criterion for markov decision processes. Management Science, 23(1):43–49.

Jarvik, M. E. (1951). Probability learning and a negative recency effect in the serial anticipation of alternative symbols. Journal of Experimental Psychology, 41(4):291–297.

Jauernig, J., Uhl, M., and Walkowitz, G. (2022). People prefer moral discretion to algorithms: Algorithm aversion beyond intransparency. Philosophy & Technology, 35(2).

Johnson, D. D. and Tierney, D. (2011). The rubicon theory of war: How the path to conflict reaches the point of no return. International Security, 36(1):7–40.

Johnson, E. J. and Goldstein, D. (2003). Do defaults save lives? Science, 302(5649):1338–1339.

Johnson, T. J., Feigenbaum, R., and Weiby, M. (1964). Some determinants and consequences of the teacher's perception of causation. Journal of Educational Psychology, 55(5):237–246.

Jones, B. D. (1999). Bounded rationality. Annual Review of Political Science, 2:297–321.

Jussupow, E., Benbasat, I., and Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. 28th European Conference on Information Systems - Liberty, Equality, and Fraternity in a Digitizing World, ECIS 2020, Marrakech, Morocco, June 15-17, 2020: Proceedings.

Jussupow, E., Benbasat, I., and Heinzl, A. (2024). An integrative perspective on algorithm aversion and appreciation in decision-making. MIS Quarterly, 48:1575–1590.

Kachelmeier, S. J. and Shehata, M. (1992). Examining Risk Preferences Under High Monetary Incentives: Experimental Evidencefrom the People's Republic of China. The American Economic Review, 82(5):1120–1141.

Kahneman, D. (2003). Maps of bounded rationality. The Armerican Economic Review, 93:1449–1475.

Kahneman, D. and Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. Management Science, 39(1):17–31.

Kahneman, D. and Tversky, A. (1972). Subjective probability: A judgment of representativeness. Cogntive Psychology, 8:430–454.

Kahneman, D. and Tversky, A. (1973). On the psychology of prediction. Psychological Review, 80(4):237–251.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47(2):263–292.

Kahneman, D. and Tversky, A. (1996). On the reality of cognitive illusions. Psychological Review, 103:582–591.

Kaiser, G. (2025). Anzahl der Visits von chatgpt.com von Juli 2024 bis Mai 2025. Technical report, Statista.

Kajackaite, A. and Gneezy, U. (2017). Incentives and cheating. Games and Economic Behavior, 102:433–444.

Kamenica, E. (2012). Behavioral economics and psychology of incentives. Annual Review of Economics, 4(1):427–452.

Karmaker, S. K., Hassan, M. M., Smith, M. J., Xu, L., Zhai, C., and Veeramachaneni, K. (2020). Automl to date and beyond: Challenges and opportunities. ACM Computing Surveys, 54(8):1–36.

Kasneci, G., Van Gael, J., Herbrich, R., and Graepel, T. (2010). Bayesian knowledge corroboration with logical rules and user feedback. In J.L., B., F., B., A., G., and M., S., editors, Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2010., pages 1–18.

Kawaguchi, K. (2021). When Will Workers Follow an Algorithm? A Field Experiment with a Retail Business. Management Science, 67(3):1670–1695.

Kayande, U., Bruyn, A. D., Lilien, G. L., Rangaswamy, A., and van Bruggen, G. H. (2009). How Incorporating Feedback Mechanisms in a DSS Affects DSS Evaluations. Information Systems Research, 20(4):527–546.

Khalifa, M. and Albadawy, M. (2024). Using artificial intelligence in academic writing and research: An essential productivity tool. Computer Methods and Programs in Biomedicine Update, 5:100145.

Khalmetski, K. and Sliwka, D. (2019). Disguising lies - Image concerns and partial lying in cheating games. American Economic Journal: Microeconomics, 11(4):79–110.

Kleinberg, J., Lakkaraju, H., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. Quarterly Journal of Economics, 133(1):237–293.

Kleinmuntz, D. N. (1985). Cognitive heuristics and feedback in a dynamic decision environment. Management Science, 31(6):680–702.

Klingbeil, A., Grützner, C., and Schreck, P. (2024). Trust and reliance on AI - An experimental study on the extent and costs of overreliance on AI. Computers in Human Behavior, 160:108352.

Koellinger, P. and Treffers, T. (2015). Joy leads to overconfidence, and a simple countermeasure. PLoS ONE, 10(12):e0143263.

Kolling, A., Walker, P., Chakraborty, N., Sycara, K., and Lewis, M. (2016). Human interaction with robot swarms: A survey. IEEE Transactions on Human-Machine Systems, 46(1):9–26.

Komperla, R. C. A. (2021). AI-enhanced claims processing: Streamlining insurance operations. Journal of Research Administration, 3(2):95–106.

Komperla, R. C. A. (2023). How can AI help in fraudulent claim identification. Journal of Research Administration, 5:1539–1590.

Konow, J. (2003). Which Is the Fairest One of All? A Positive Analysis of Justice Theories. Journal of Economic Literature, 41(4):1188–1239.

Korinek, A. (2023). Generative AI for Economic Research: Use Cases and Implications for Economists. Journal of Economic Literature, 61:1281–1317.

Kouziokasa, G. N. (2017). The application of artificial intelligence in public administration for forecasting high crime risk transportation areas in urban environment. Transportation Research Procedia, 24:467–473.

Kraft, P. S., Günther, C., Kammerlander, N. H., and Lampe, J. (2022). Overconfidence and entrepreneurship: A meta-analysis of different types of overconfidence in the entrepreneurial process. Journal of Business Venturing, 37(4):106207.

Krahnen, J. P., Ockenfels, P., and Wilde, C. (2014). Measuring ambiguity aversion: A systematic experimental approach. SAFE Working Paper No. 55.

Krawczyk, M. (2012). Incentives and timing in relative performance judgments: A field experiment. Journal of Economic Psychology, 33:1240–1246.

Kruger, J. (1999). Lake wobegon be gone! The 'below-average effect' and the egocentric nature of comparative ability judgments. Journal of Personality and Social Psychology, 77(2):221–232.

Kruger, J., Windschitl, P. D., Burrus, J., Fessel, F., and Chambers, J. R. (2008). The rational side of egocentrism in social comparisons. Journal of Experimental Social Psychology, 44:220–232.

Krügel, S., Ostermaier, A., and Uhl, M. (2022). Zombies in the loop? Humans trust untrustworthy ai-advisors for ethical decisions. Philosophy & Technology, 35(17).

Köbis, N., Bonnefon, J.-F., and Rahwan, I. (2021). Bad machines corrupt good morals. Nature human behavior, 5(6):79–94.

Lam, C.-P. and Sastry, S. S. (2014). A POMDP framework for human-in-the-loop system. In Proc. IEEE Conference on Decision and Control, pages 6031–6036.

Lamal, P. A. (1990). On the importance of replication. Journal of Social Behavior and Personality, 5(4):31–35.

Landier, A. and Thesmar, D. (2003). Financial contracting with optimistic entrepreneurs. The Review of Financial Studies, 22(1):117–150.

Langer, E. J. (1975). The illusion of control. Journal of Personality and Social Psychology, 32(2):311–328.

Larkin, I. and Leider, S. (2012). Does commitment or feedback influence myopic loss aversion? An experimental analysis. American Economic Journal: Microeconomics, 4(2):184–214.

Larrick, R. P., Burson, K. A., and Soll, J. B. (2007). Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not). Organizational Behavior and Human Decision Processes, 102(1):76–94.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data & Society, 5(1):1–16.

Leib, M., Köbis, N. C., Rilke, R. M., Hagens, M., and Irlenbusch, B. (2024). Corrupted by algorithms? How AI-generated and human-written advice shape (dis)honesty. The Economic Journal, 134:766–784.

Levinson, J., Askeland, J., Becker, J., Dolson, J., Held, D., Kammel, S., Kolter, J. Z., Langer, D., Pink, O., Pratt, V., Sokolsky, M., Stanek, G., Stavens, D., Teichman, A., Werling, M., and Thrun, S. (2008). Towards fully autonomous driving: Systems and algorithms. 2011 IEEE Intelligent Vehicles Symposium (IV), Baden-Baden, Germany, June 5-9, 2011, pages 163–168.

Li, J., Liu, W., and Zhang, J. (2025). Automating financial audits with random forests and real-time stream processing: A case study on efficiency and risk detection. Informatica, 49:1–20.

Libby, R. and Rennekamp, K. (2012). Self-serving attribution bias, overconfidence, and the issuance of management forecasts. Journal of Accounting Research, 50(1):197–231.

Lichtenstein, S. and Fischhoff, B. (1977). Do those who know more also know more about how much they know? Organizational Behavior and Human Performance, 20(2):159–183.

Litterscheidt, R. and Streich, D. J. (2020). Financial education and digital asset management: what's in the black box? Journal of Behavioral and Experimental Economics, 87:101573.

Liu, M., Brynjolfsson, E., and Dowlatabadi, J. (2021). Do digital platforms reduce moral hazard? the case of uber and taxis. Management Science, 67(8):4665–4685.

Loewenstein, G., Weber, E. U., Hsee, C. K., and Welch, N. (2001). Risk as feelings. Psychological Bulletin, 127(2):267–286.

Loewenstein, G. F., Thomspon, L., and Bazerman, M. H. (1989). Social utility and decision making in interpersonal contexts. Journal of Personality and Social Psychology, 57:426–441.

Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151:90–103.

Longoni, C., Bonezzi, A., and Morewedge, C. K. (2019). Resistance to medical artificial intelligence. Journal of Consumer Research, 46(4):629–650.

Madhavan, P. and Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: an integrative review. Theoretical Issues in Ergonomics Science, 8(4):277–301.

Mahar, H. (2003). Why are there so few prenuptial agreements? Harvard Law School John M. Olin Center for Law, Economics and Business. Discussion paper No. 436.

Mahmud, H., Islam, A. N., Ahmed, S. I., and Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. Technological Forecasting and Social Change, 175:121390.

Majumdar, A., Singh, S., Mandlekar, A., and Pavone, M. (2017). Risk-sensitive inverse reinforcement learning via coherent risk models. Conference Paper: Robotics: Science and Systems.

Malmendier, U. and Nagel, S. (2011). Depression babies: Do macroeconomic experiences affect risk taking? The Quarterly Journal of Economics, 126(1):373–416.

Malmendier, U. and Tate, G. (2005). CEO overconfidence and corporate investment. The Journal of Finance, 60(6):2661–2700.

Malmendier, U. and Tate, G. (2008). Who makes acquisitions? CEO overconfidence and the market's reaction. The Journal of Financial Economics, 89(1):20–43.

March, J. G. (1978). Bounded rationality, ambiguity, and the engineering of choice. The Bell Journal of Economics, 9:587–608.

Martinelli, C., Parker, S. W., Pérez-Gea, A. C., and Rodrigo, R. (2018). Cheating and incentives. American Economic Journal: Economic Policy, 10:298–325.

Mayraz, G. (2011). Wishful thinking. CEP Discussion Papers.

Mazar, N., Amir, O., and Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. Journal of Marketing Research, 45(6):633–644.

Merkle, C. and Weber, M. (2011). True overconfidence: The inability of rational information processing to account for apparent overconfidence. Organizational Behavior and Human Decision Processes, 116:262–271.

Miller, J. B. and Sanjurjo, A. (2018a). How experience confirms the gambler's fallacy when sample size is neglected. Working paper, pages 1–12.

Miller, J. B. and Sanjurjo, A. (2018b). Surprised by the hot hand fallacy? a truth in the law of small numbers. Econometrica, 86(6):2019–2047.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267:1–38.

Mittelstadt, B. D., Allo1, P., Taddeo, M., Wachter, S., and Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, pages 1–21.

Mol, J. M., van der Heijden, E. C. M., and Potters, J. J. M. (2020). (Not) alone in the world: Cheating in the presence of a virtual observer. Experimental Economics, 23:961–978.

Mollerstrom, J., Reme, B. A., and Sørensen, E. Ø. (2015). Luck, choice and responsibility: An experimental study of fairness views. Journal of Public Economics, 131:33–40.

Moore, D. A. and Cain, D. M. (2007). Overconfidence and underconfidence: When and why people underestimate (and overestimate) the competition. Organizational Behavior and Human Decision Processes, 103(2):197–213.

Moore, D. A. and Healy, P. J. (2008). The trouble with overconfidence. Psychological Review, 115(2):502–517.

Moore, D. A. and Schatz, D. (2017). The three faces of overconfidence. Social and Personality Psychology Compass, 11(8):1–12.

Moore, D. A. and Small, D. A. (2007). Error and bias in comparative judgment: On being both better and worse than we think we are. Journal of Personality and Social Psychology, 92(6):972–989.

Moore, E. and Eckel, C. (2003). Measuring ambiguity aversion. Unpublished manuscript, Department of Economics, Virginia Tech.

Mousavi, S. and Gigerenzer, G. (2014). Risk, uncertainty, and heuristics. Journal of Business Research, 67:1671–1678.

Narayanan, S. and Georgiou, P. G. (2013). Behavioral signal processing: Deriving human behavioral informatics from speech and language. Proceedings of the IEEE, 101(5):1203–1233.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. European Journal of Social Psychology, 15(3):263–280.

Niszczota, P. and Kaszás, D. (2020). Robo-investment aversion. PLoS ONE, 15(9):e0239277.

Norton, M. I., Mochon, D., and Ariely, D. (2012). The IKEA effect: When labor leads to love. Journal of Consumer Psychology, 22(3):453–460.

Odean, T. (1998). Volume, volatility, price, and profit when all traders are above average. The Journal of Finance, 53(6):1887–1934.

Odean, T. (1999). Do investors trade too much? The American Economic Review, 89(5):1279–1298.

Önkal, D., Goodwin, P., Thomson, M., Gönül, S., and Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. Journal of Behavioral Decision Making, 22(4):390–409.

Owens, D., Grossman, Z., and Fackler, R. (2014). The control premium: A preference for payoff autonomy. American Economic Journal: Microeconomics, 6(4):138–161.

Palan, S. and Schitter, C. (2018). Prolific.ac – A subject pool for online experiments. Journal of Behavioral and Experimental Finance, 17:22–27.

Parasuraman, R. and Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39:230–253.

Park, Y. J. and Santos-Pinto, L. (2010). Overconfidence in tournaments: Evidence from the field. Theory and Decision, 69:143–166.

Pearl, J. (2003). Causality: Models, reasoning, and inference. Econometric Theory, 19(4):675–682.

Peer, E., Acquisti, A., and Shalvi, S. (2014). "I cheated, but only a little": Partial confessions to unethical behavior. Journal of Personality and Social Psychology, 106(2):202–217.

Peer, E., Brandimarte, L., Samat, S., and Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. Journal of Experimental Social Psychology, 70:153–163.

Petisca, S., Leite, I., Paiva, A., and Esteves, F. (2022). Human dishonesty in the presence of a robot: The effects of situation awareness. International Journal of Social Robotics, 14:1211–1222.

Pittarello, A., Leib, M., Gordon-Hecker, T., and Shalvi, S. (2015). Justifications shape ethical blind spots. Psychological science, 26(6):794–804.

Prahl, A. and Swol, L. V. (2017). Understanding algorithm aversion: When is advice from automation discounted? Journal of Forecasting, 36(6):691–702.

Pratt, J. W. (1964). Risk aversion in the small and in the large. Econometrica, 32(1/2):122–136.

Presson, P. K. and Benassi, V. A. (1996). Illusion of control: A meta-analytic review. Journal of Social Behavior and Personality, 11(3):493–510.

Prieto-Gutierrez, J.-J., Segado-Boj, F., and França, F. D. S. (2023). Artificial intelligence in social science: A study based on bibliometrics analysis. Human Technology, 19:149–162.

Prissé, B. and Jorrat, D. (2022). Lab vs online experiments: No differences. Journal of Behavioral and Experimental Economics, 100:101910.

Proeger, T. and Meub, L. (2014). Overconfidence as a social bias: Experimental evidence. Economic Letters, 122(2):203–207.

Puterman, M. L. (1994). Markov Decision Processes. Wiley-Interscience, Hoboken, N.J.

Rabin, M. (1993). Incorporating fairness into game theory and economics. American Economic Review, 83(5):1281–1302.

Rabin, M. and Vayanos, D. (2009). The gambler's and hot-hand fallacies: Theory and applications. Review of Economic Studies, 77(2):730—778.

Radzevick, J. R. and Moore, D. A. (2011). Competing to be certain (but wrong): Market dynamics and excessive confidence in judgment. Management Science, 57(1):93–106.

Rahwan, Z., Hauser, O. P., Kochanowskae, E., and Fasoloa, B. (2018). High stakes: A little more cheating, a lot less charity. Journal of Economic Behavior & Organization, 152:276–295.

Read, D., Loewenstein, G., and Rabin, M. (1999). Choice bracketing. Journal of Risk and Uncertainty, 19(1-3):171–197.

Renier, L. A., Mast, M. S., and Bekbergenova, A. (2021). To err is human, not algorithmic – robust reactions to erring algorithms. Computers in Human Behavior, 124(106879):1–12.

Rey-Biel, P., Sheremeta, R., and Uler, N. (2015). When income depends on performance and luck: The effects of culture and information on giving. Research in Experimental Economics, 20:167–203.

Rosenthal, R. (1990). Replication in behavioral research. Journal of Social Behavior and Personality, 5(4):1–30.

Roth, A. E. (1988). Laboratory experimentation in economics: A methodological overview. The Economic Journal, 98(393):974–1031.

Roth, A. E. and Peranson, E. (1999). The redesign of the matching market for american physicians: Some engineering aspects of economic design. The American Economic Review, 89(4):748–780.

Samuelson, P. A. and Nordhaus, W. (1985). Principles of Economics. McGraw-Hill, 12 edition.

Sandoval, E. B., Brandstatter, J., Yalcin, U., and Bartneck, C. (2020). Robot likeability and reciprocity in human robot interaction. International Journal of Social Robotics, 13(10):851–862.

Sandroni, A. and Squintani, F. (2007). Overconfidence, insurance, and paternalism. The American Economic Review, 97(5):1994–2004.

Sandroni, A. and Squintani, F. (2013). Overconfidence and asymmetric information: The case of insurance. Journal of Economic Behavior & Organization, 93:149–165.

Santos-Pinto, L. and de la Rosa, L. E. (2020). Overconfidence in labor markets. In Zimmermann, K. F., editor, Handbook of Labor, Human Resources and Population Economics, pages 1–42. Springer Nature.

Sarstedt, M., Marko, Neubert, D., and Barth, K. (2016). The IKEA Effect. A Conceptual Replication. Journal of Marketing Behavior, 2:307–312.

Schaefer, P. S., Williams, C. C., Goodie, A. S., and Campbell, W. K. (2004). Overconfidence and the Big Five. Journal of Economic Research in Personality, 38:473–480.

Scharowski, N., Perrig, S. A. C., Svab, M., Opwis, K., and Brühlmann, F. (2023). Exploring the effects of human-centered AI explanations on trust and reliance. Frontiers in Computer Science, 5:1151150.

Scheinkman, J. A. and Xiong, W. (2003). Overconfidence and speculative bubbles. Journal of Political Economy, 111(6):1183–1220.

Schirner, G., Erdogmus, D., Chowdhury, K., and Padir, T. (2013). The Future of Human-in-the-Loop Cyber-Physical Systems. Computer, 46(1):36–45.

Schlund, R. and Zitek, E. M. (2024). Algorithmic versus human surveillance leads to lower perceptions of autonomy and increased resistance. Communications Psychology, 2(1):53.

Schmitt, A., Wambsganss, T., Söllner, M., and Janson, A. (2021). Towards a Trust Reliance Paradox? Exploring the Gap Between Perceived Trust in and Reliance on Algorithmic Advice. International Conference on Information Systems (ICIS).

Schneider, S. M. and Schupp, J. (2014). Individual Differences in Social Comparison and its Consequences for Life Satisfaction: Introducing a Short Scale of the Iowa–Netherlands Comparison Orientation Measure. Social Indicators Research, 115:767–789.

Schubert, T. W. and Otten, S. (2002). Overlap of self, ingroup, and outgroup: Pictorial measures of self-categorization. Self and Identity, 1(4):353–376.

Sedlmeier, P. and Gigerenzer, G. (2001). Teaching bayesian reasoning in less than two hours. Journal of Experimental Psychology, 130(3):380–400.

Selten, R. (1998). Features of experimentally observed bounded rationality. European Economic Review, 42:413–436.

Selten, R. (2001). What is bounded rationality? In Gigerenzer, G. and Selten, R., editors, Bounded Rationality: The Adaptive Toolbox, volume 1, pages 13–36. Cambridge, MA: The MIT Press.

Semrush.com (2025). Explore the world's most visited websites. https://www.semrush.com/trending-websites/global/all [accessed 01.07.2025].

Shalvi, S., Dana, J., Handgraaf, M. J., and Dreu, C. K. D. (2011). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. Organizational Behavior and Human Decision Processes, 115(2):181–190.

Shalvi, S., Levine, E., Thielmann, I., Jayawickreme, E., van Rooij, B., Teodorescu, K., Schurr, A., Furr, M., Aglioti, S. M., Zettler, I., Cohen, T., Pittarello, A., Barkan, R., Köbis, N., Leib, M., Mitkidis, P., Schulz, J. F., Dimant, E., van Kleef, G., Ścigała, K. A., Rilke, R. M., Ayal, S., Beersma, B., Plonsky, O., Hilbig, B. E., Weisel, O., Butera, F., Feldman, Y., Verschuere, B., Zanetti, C., Hochman, G., Kret, M., Peer, E., Capraro, V., Dorrough, A. R., Speer, S., and Ritov, I. (2025). The science of honesty: A review and research agenda. Forthcoming.

Sharan, N. N. and Romano, D. M. (2020). The effects of personality and locus of control on trust in humans versus artificial intelligence. Helyion 6, e04572.

Shiller, R. J. (2003). From efficient markets theory to behavioral finance. Journal of Economic Perspectives, 17:83–104.

Shin, D. (2020). User perceptions of algorithmic decisions in the personalized AI system: Perceptual evaluation of fairness, accountability, transparency, and explainability. Journal of Broadcasting & Electronic Media, 64(4):541–565.

Simon, H. A. (1955). A behavioral model of rational choice. The Quarterly Journal of Economics, 69:99–118.

Simon, H. A. (1972). Theories of bounded rationality. In Decision and Organization. North-Holland.

Simon, H. A. (1978). Rationality as process and as product of thought. The American Economic Review, 68:1–16.

Simon, H. A. (1986). Rationality in psychology and economics. Journal of Business, 59:S209–S224.

Simon, H. A. (1989). The scientist as problem solver. In Complex Information Processing: The Impact of Herbert A. Simon, pages 375–398. Routledge.

Simon, M., Houghton, S. M., and Aquino, K. (2000). Cognitive biases, risk perception, and venture formation. Journal of Business Venturing, 15(2):113–134.

Singh, V. and Joshi, K. (2022). Automated Machine Learning (AutoML): an overview of opportunities for application and research. Journal of Information Technology Case and Application Research, 24:1–11.

Slovic, P., Fischhoff, B., and Lichtenstein, S. (1978). Accident probabilities and seat belt usage: A psychological perspective. Accident Analysis & Prevention, 10(4):281–285.

Slovic, P., Fischhoff, B., and Lichtenstein, S. (1982). Why study risk perception? Risk Analysis, 2(2):83–93.

Smith., N. C. (1970). Replication studies: A neglected aspect of psychological research. American Psychologist, 25(10):970–975.

Smith, V. L. (1976). Experimental economics: Induced value theory. The American Economic Review, 66(2):274–279.

Starmer, C. and Sugden, R. (1991). Does the random-lottery incentive system elicit true preferences? An experimental investigation. Journal of Economic Behavior & Organization, 81:971–978.

Statman, M., Thorley, S., and Vorkink, K. (2006). Investor overconfidence and trading volume. The Review of Financial Studies, 19(4):1531–1565.

Strang, L. and Schaube, S. (2025). (Not) everyone can be a winner – The role of payoff interdependence for redistribution. Journal of Public Economics, 243:105320.

Subbotin, V. (1996). Outcome feedback effects on under- and overconfident judgments (general knowledge tasks). Organizational Behavior and Human Decision Processes, 66(3):268–276.

Suetens, S., Galbo-Jørgensen, C. B., and Tyran, J.-R. (2016). Predicting lotto numbers: A natural experiment on the gambler's fallacy and the hot-hand fallacy. Journal of the European Economic Association, 14:584–607.

Sutherland, S. C., Harteveld, C., and Young, M. E. (2016). Effects of the advisor and environment on requesting and complying with automated advice. ACM Transactions on Interactive Intelligent Systems, 6:1–36.

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? Acta Psychologica, 47(2):143–148.

Tanelli, M., Toledo-Moreo, R., and Stanley, L. M. (2017). Guest editorial: Holistic approaches for human–vehicle systems: Combining models, interactions, and control. IEEE Trans. Human-Machine Systems, 47(5):609–613.

Tao, R., Su, C.-W., Xiao, Y., Dai, K., and Khalid, F. (2021). Robo advisors, algorithmic trading and investment management: Wonders of fourth industrial revolution in financial markets. Technological Forecasting and Social Change, 163:120421.

Taylor, S. E. and Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. Psychological Bulletin, 103(2):193–210.

Tenney, E. R., Meikle, N. L., Hunsaker, D., Moore, D. A., and Anderson, C. (2019). Is overconfidence a social liability? The effect of verbal versus nonverbal expressions of confidence. Journal of Personality and Social Psychology, 116(3):396–415.

Thaler, R. H. and Benartzi, S. (2004). Save more tomorrow: Using behavioral economics to increase employee saving. Journal of Political Economy, 112(S1):164–187.

Thaler, R. H. and Sunstein, C. R. (2008). Nudge: Improving Decisions About Health, Wealth, and Happiness. Yale University Press.

Todd, P. M. and Gigerenzer, G. (2003). Bounding rationality to the world. Journal of Economic Psychology, 24:143–165.

Törngren, G. and Montgomery, H. (2004). Worse Than Chance? Performance and Confidence Among Professionals and Laypeople in the Stock Market. The Journal of Behavioral Finance, 5(3):148–153.

Townsend, D. M., Busenitz, L. W., and Arthurs, J. D. (2010). To start or not to start: Outcome and ability expectations in the decision to start a new venture. Journal of Business Venturing, 25(2):192–202.

Trinugroho, I. and Sembel, R. (2011). Overconfidence and excessive trading behavior: An experimental study. International Journal of Business and Management, 6(7):147–152.

Tsai, C. I., Klayman, J., and Hastie, R. (2008). Effects of amount of information on judgment accuracy and confidence. Organizational Behavior and Human Decision Processes, 107(2):97–105.

Tschider, C. A. (2020). Beyond the 'black box'. Denver Law Review, 98(3):683–723.

Turney, P. D., Littman, M. L., Bigham, J., and Shnayder, V. (2003). Combining independent modules to solve multiple-choice synonym and analogy problems. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03), pages 482–489.

Tversky, A. and Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. Cognitive Psychology, 5:207–232.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185(4157):1124–1131.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. Science, 211(4481):453–458.

Tversky, A. and Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. Psychological Review, 90:293–315.

Ullman, D., Leite, I., Phillips, J., Kim-Cohen, J., and Scassellati, B. (2014). Smart human, smarter robot: How cheating affects perceptions of social agency. Proceedings of the Annual Meeting of the Cognitive Science Society, 36:2996–3001.

Urbig, D., Stauf, J., and Weitzel, U. (2009). What is your level of overconfidence? A strictly incentive compatible measurement of absolute and relative overconfidence. Working Papers 09-20, Utrecht School of Economics.

Vaccaro, M., Almaatouq, A., and Malone, T. (2024). When combinations of humans and ai are useful: A systematic review and meta-analysis. Nature Human Behaviour, 8:2293–2303.

van Hippel, W. and Trivers, R. (2011). The evolution and psychology of self-deception. Behavioral and Brain Sciences, 34(1):1 –56.

van Otterlo, M. and Wiering, M. (2012). Reinforcement learning and markov decision processes. In Wiering, M. and van Otterlo, M., editors, Reinforcement Learning, volume 12, chapter 1, pages 3–42. Berlin: Springer.

van Overloop, P. J., Maestre, J. M., Sadowska, A. D., Camacho, E. F., and de Schutter, B. (2015). Human-in-the-loop model predictive control of an irrigation canal. IEEE Control Systems Magazine, 35(4):19–29.

Varian, H. R. (2006). Intermediate Microeconomics: A Modern Approach. New York, NY: W. W. Norton & Company, 8 edition.

Vempaty, A., Kailkhura, B., and Varshney, P. K. (2018). Human-machine inference networks for smart decision making: Opportunities and challenges. In IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), pages 6961–6965.

von Neumann, J. and Morgenstern, O. (1944). Theory of Games and Economic Behavior. Princeton University Press, 2 edition.

Vörös, Z. (2024). Effect of the different forms of overconfidence on venture creation: Overestimation, overplacement and overprecision. Journal of Management & Organization, 30:304–317.

Wagner-Menghin, M. (2004). Lexikon-wissen-test (LEWITE). Unpublished Dissertation, University of Vienna.

Wang, F. A. (2001). Overconfidence, investor sentiment, and evolution. Journal of Financial Intermediation, 10(2):138–170.

Wang, Y., Wang, Z., and Li, J. (2024). Does algorithmic control facilitate platform workers' deviant behavior toward customers? The ego depletion perspective. Computers in Human Behavior, 156:108242.

Weber, E. U. (2006). Experience-based and description-based perceptions of long-term risk: Why global warming does not scare us (yet). Climate Change, 77(1-2):103–120.

Weimann, J. and Brosig-Koch, J. (2019). Einführung in die exerimentelle Wirtschaftsforschung. Springer Gabler. https://doi.org/10.1007/978-3-642-32765-0.

Weinstein, N. D. (1980). Unrealistic optimism about future life events. Journal of Personality and Social Psychology, 39(5):806–820.

World Health Organization (2025). WHO launches new membership of expert group for behavioural sciences. https://www.who.int/news/item/08-04-2025-who-launches-new-membership-of-expert-group-for-behavioural-sciences [accessed 07.07.2025].

Yeomans, M., Shah, A., Mullainathan, S., and Kleinberg, J. (2019). Making sense of recommendations. Information Systems Research, 32(4):403–414.

Yost, A. B., Behrend, T. S., Howardson, G., Darrow, J. B., and Jensen, J. M. (2019). Reactance to electronic surveillance: A test of antecedents and outcomes. Journal of Business and Psychology, 34:71–86.

Young, S. N. and Peschel, J. M. (2020). Review of human–machine interfaces for small unmanned systems with robotic manipulators. IEEE Transactions on Human-Machine Systems, 50(2):131–143.

Zacharakis, A. L. and Shepherd, D. A. (2001). The nature of information and overconfidence on venture capitalists' decision making. Journal of Business Venturing, 16(4):311–332.

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J., and Li, Y. (2021). A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. Complexity, pages 1–18.

Zöller, M.-A., Titov, W., Schlegel, T., and Huber, M. (2022). XAutoML: A Visual Analytics Tool for Establishing Trust in Automated Machine Learning. https://arxiv.org/abs/2202.11954.

APPENDIX

Supplementary Materials to Chapter 3

A.1 Derivation of Payoff Utility Function

Table 29: Derivation of payoff-maximizing decision strategy

		Report					
		1	2	3	4	5	6
Number drawn							
	P(audit)	0.1	0.2	0.3	0.4	0.5	0.6
	P(punishment|audit)	0	0.5	0.667	0.75	0.8	0.833
	P(punishment)	0	0.1	0.2	0.3	0.4	0.5
1	P(cheating successful)	0	0.9	0.8	0.7	0.6	0.5
	E(payoff)	15	27.75	37.5	44.25	48	**48.75**
	U(cheating)		12.75	22.5	29.25	33	33.75
	U'(cheating)		12.75	9.75	6.75	3.75	0.75
	P(audit)		0.2	0.3	0.4	0.5	0.6
	P(punishment|audit)		0	0.333	0.5	0.6	0.667
	P(punishment)		0	0.1	0.2	0.3	0.4
2	P(cheating successful)			0.9	0.8	0.7	0.6
	E(payoff)			42.00	51.00	57.00	**60.00**
	U(cheating)			12	21	27	30
	U'(cheating)			12	9	6	3
	P(audit)			0.3	0.4	0.5	0.6
	P(punishment|audit)				0.25	0.4	0.5
	P(punishment)				0.1	0.2	0.3
3	P(cheating successful)				0.9	0.8	0.7
	E(payoff)				56.25	64.50	**69.75**
	U(cheating)				11.25	19.5	24.75
	U'(cheating)				11.25	8.25	5.25
	P(audit)				0.4	0.5	0.6
	P(punishment|audit)				0	0.2	0.333
	P(punishment)				0	0.1	0.2
4	P(cheating successful)				0	0.9	0.8
	E(payoff)				60	70.50	**78.00**
	U(cheating)					10.50	18
	U'(cheating)					10.50	7.50
	P(audit)					0.5	0.6
	P(punishment|audit)					0	0.167
	P(punishment)					0	0.1
5	P(cheating successful)					0	0.9
	E(payoff)					75	**84.75**
	U(cheating)						9.75
	U'(cheating)						9.75
	P(audit)						0.6
	P(punishment|audit)						0
	P(punishment)						0
6	P(cheating successful)						0
	E(payoff)						90
	U(cheating)						
	U'(cheating)						

Note: Abbreviations in table denote the following: $P(x)$: probability; $P(x_1|x_2)$: conditional probability; $E(x)$: expected value; $U(x)$: utility; $U'(x)$: marginal utility.

A.2 Experiment Instructions

[Translated from German]

General information

- For your participation in the experiment you will receive a fixed payoff - including the show-up fee - of €7.50.
- Additionally, you can receive a prize of up to €90 in a prize draw.

Procedure

- You will first be asked to answer some comprehension questions about these instructions.
- The experiment starts as soon as all participants have read the instructions and answered the comprehension questions correctly.

Drawing a card for the prize draw

- First you draw a card from an urn. Please take the drawn card, keep it safe and do not show it to anyone.
- There are 100 cards in the urn at the beginning. On each card there is a number from 1 - 6.

Input your number

- You will then be asked to report your number via the input field on your screen and confirm the entry.
- The number you report determines the amount of your potential additional prize. The additional prize is calculated by multiplying your reported number by €15. Possible win amounts would be accordingly:

Reported number	**Additional prize**
1	€15.00
2	€30.00
3	€45.00
4	€60.00
5	€75.00
6	€90.00

Questionnaire

- You will then be asked to complete a multi-part questionnaire.
- All answers in the questionnaire remain completely anonymous and have **no effect on your chance of winning the prize draw**.

Prize draw / Payout

Prize draw

- After all participants have completed the questionnaire, one participant will be drawn at random to receive the additional prize.
- The draw will take place base on the cabin numbers.
- All participants have the same chance to receive the additional prize regardless of their reported number.

Payout

- After the winner has been determined, all participants who do not receive an additional prize will be paid first. You will be called by your cabin number and receive the fixed payment.
- The payout of the draw prize, as well as the potential verification, will take place after all other participants have left the lab.

The payout process for the additional prize consists of the following steps:
[Non-blackbox treatments:]

Lottery 1: Decision on verification of your card

- An experimenter [An algorithm] decides on the check, i.e. whether your reported number is compared with your card.
- The experimenter [algorithm] randomly draws a number between 1 and 10 from a lottery pot (all numbers are equally likely).
  - If the **number drawn** by the experimenter [algorithm] is **higher than the number you reported**, you don't have to reveal your card and you receive your designated payoff - your Reported Number x €15 - immediately. In this case, the experiment is over.

- If the number drawn by the experimenter [algorithm] is lower than or equal to the number you reported, it is checked whether the number on your card matches the number you reported.

Depending on the outcome of Lottery 1: Card check

- If the number you report matches the number on your card, you will receive your payoff - your Reported Number x €15 - in full. In this case, the experiment is over.
- If the number you reported does not match the number on your card, Lottery 2 is played. This will decide whether your payoff will be reduced.

Depending on the outcome of the check: Lottery 2 & Potential adjustment of the payoff

- A lottery pot is filled with numbers from 1 to your reported number (in integer steps). The experimenter then randomly draws a number [The algorithm randomly draws a number that can take values from 1 to your reported number (in integer steps)].
  - If the number drawn by the experimenter [algorithm] is lower than or equal to the number on your card, you will receive the full payoff, i.e. your Reported Number x €15.
  - If the number drawn by the experimenter [algorithm] is higher than the number on your card, you will receive a reduced payout depending on the number on your card - Number on Card x €7.50. Accordingly, possible winning amounts would be:

Number on card	**Additional prize**
1	€7.50
2	€15.00
3	€22.50
4	€30.00
5	€37.50
6	€45.00

- This means, you cannot go away empty-handed if you are drawn for the additional prize.
- In both cases the experiment is finished afterwards.
- To summarize: The experimenter [algorithm] performs at least 1 and max. 2 lotteries during the payout process.

[Blackbox treatments:]

Decision 1: Decision on verification of your card

- An experimenter [An algorithm] decides on the check, i.e. whether your reported number is compared with your card.
- If the experimenter [algorithm] decides not to inspect your card, you will receive your payoff - your Reported Number x €15 - immediately. In this case, the experiment is over.

Depending on the outcome of Decision 1: Card check

- If the number you report matches the number on your card, you will receive your payoff - your Reported Number x €15 - in full. In this case, the experiment is over.
- If the number you reported does not match the number on your card, the experimenter [the algorithm] will decide whether your payoff will be reduced.

Depending on the outcome of the check: Decision 2 & Potential adjustment of the payoff

- If the experimenter [the algorithm] decides that your payoff will not be reduced, you will receive the full payoff, i.e. your Reported Number x €15. If the experimenter [the algorithm] decides that your payoff will be reduced, you will receive a reduced payout depending on the number on your card - Number on Card x €7.50. Accordingly, possible winning amounts would be:

Number on card	Additional prize
1	€7.50
2	€15.00
3	€22.50
4	€30.00
5	€37.50
6	€45.00

- This means, you cannot go away empty-handed if you are drawn for the additional prize.
- In both cases the experiment is finished afterwards.
- To summarize: The experimenter [algorithm] makes at least 1 and max. 2 decisions during the payout process.

Additional remarks

- No communication is allowed during the experiment.

- All decisions you make during this experiment will be completely anonymous. None of the other participants will learn of your identity, the decisions you make, or the payoff you receive. The data will be analyzed for scientific purposes only.

Good luck and thank you for participating in this experiment!

Comprehension questions

Please answer the following questions.

[Questions asked in all treatments, translated form German, 'X' indicates correct answer]

Who will receive a bonus payment?

- ○ All participants will receive a bonus payment.
- ○ One half of the participants will receive a bonus payment.
- ○ One participant will receive a bonus payment. X

Which statement regarding the payoff process is correct?

- ○ The payout of the bonus payment takes place in camera. X
- ○ The amount of the bonus payment exclusively depends on the reported number.
- ○ The amount of the bonus payment is fixed.

Which payment amount total is the minimum you will receive in case you are drawn to receive the bonus payment?

- ○ €7.50 (i.e., the fixed payment)
- ○ €15.00 (i.e., the fixed payment + €7.50) X
- ○ €22.50 (i.e., the fixed payment + €15)

What happens in case your card is inspected?

- ○ You receive a new card.
- ○ You only receive the show-up fee.
- ○ Your reported number will be compared with the number on your card. X

Please answer the following questions.

[Questions asked in human treatments, translated form German, 'X' indicates correct answer]

Which statement about Lottery 1 is correct?

- ○ All participants play Lottery 1.
- ○ The drawable numbers from 1 to 10 have different probabilities.
- ○ Since a card check takes place if the number drawn in Lottery 1 is lower than or equal to the number you reported, the higher your reported number, the higher the probability of your card getting checked. X

Which statement regarding the card check is correct?

- ○ Whether a check takes place is decided by yourself.
- ○ If your reported number does not match the number on your card in a check, Lottery 2 follows. You still have a chance of receiving the full payoff (i.e., your reported number x €15). X
- ○ If your reported number does not match the number on your card in a check, you receive a reduced payoff (i.e., the number you drew x €7.50).

Which statement about Lottery 2 is correct?

- ○ If the number drawn in Lottery 2 (between 1 and the number you reported) is higher than the number on your card, your payoff is reduced (to 'number on your card' x €7.50). X
- ○ Every prize draw winner plays Lottery 2.
- ○ If the number drawn in Lottery 2 (between 1 and the number you reported) is lower than or equal to the number on your card, your payoff is reduced (to 'number on your card' x €7.50).

Please answer the following questions.

[Questions asked in machine treatments, translated form German, 'X' indicates correct answer)]

Which of the following statements regarding Decision 1 & 2 is correct?

- ○ Decision 1 is made for all participants.
- ○ Decision 2 is always made for the winner.
- ○ Decision 1 decides whether a card check will take place. X

Which of the following statements regarding the card check is correct?

- ○ Even if the number on your card does not match the number you reported, you can still receive the full payoff (i.e., your reported number x €15). X
- ○ Even if the number on your card matches the number you reported, Decision 2 follows.
- ○ If your reported number matches the number on your card, you will receive a payoff in the amount of your drawn number x €7.50.

A.3 Experiment Questionnaire

Thank you for putting in your number. You will find out whether you receive the additional prize at the end of the experiment.

In the following, we ask you to fill out our multi-part questionnaire. There is no "right" or "wrong" here. Simply answer the questions in the way that seems most appropriate to you personally. The questionnaire consists of 7 parts in total, each containing a different number of questions.

Your answers will be treated completely anonymously and will not affect your chances of winning.

Please answer the following questions.

Please recall again the verification process described for the previous decision.

Which of the following entities would you prefer to be audited by in this process?

○ a human
○ a machine (e.g. algorithm, AI, computer program, ...)

Which of the following entities do you consider to have more decision discretion?

○ a human
○ a machine (e.g. algorithm, AI, computer program, ...)

Which of the following entities do you consider more prone to making mistakes/errors?

○ a human
○ a machine (e.g. algorithm, AI, computer program, ...)

Please note: Your answers have no effect on your chance of winning the prize draw.

Please answer the following questions.

In the following questionnaire, we will ask you about your interaction with technical systems. The term "technical systems" refers to apps and other software applications, as well as entire digital devices (e.g., mobile phone, computer, TV, car navigation).

Please indicate to what extent you agree to the following statements.

Please note: Your answers have no effect on your chance of winning the prize draw.

	Completely disagree	Largely disagree	Slightly disagree	Slightly agree	Largely agree	Completely agree
I like to occupy myself in greater detail with technical systems.	○	○	○	○	○	○
I like testing the functions of new technical systems.	○	○	○	○	○	○
I predominantly deal with technical systems because I have to.	○	○	○	○	○	○
When I have a new technical system in front of me, I try it out intensively.	○	○	○	○	○	○
I enjoy spending time becoming acquainted with a new technical system.	○	○	○	○	○	○
It is enough for me that a technical system works; I don't care how or why.	○	○	○	○	○	○
I try to understand how a technical system exactly works.	○	○	○	○	○	○
It is enough for me to know the basic functions of a technical system.	○	○	○	○	○	○
I try to make full use of the capabilities of a technical system.	○	○	○	○	○	○

Please answer the following questions.

For the following statements, please indicate to what extent you consider the actions or behaviours described to be ethically problematic. Please indicate your assessment between "Definitely not problematic" and "Definitely problematic" on the following scale:

Please note: Your answers have no effect on your chance of winning the prize draw.

	Definitely unproblematic	Rather unproblematic	Not sure	Rather problematic	Definitely problematic
Taking some questionable deductions on your income tax return.	○	○	○	○	○
Having an affair with a married man/woman.	○	○	○	○	○
Passing off somebody else's work as your own.	○	○	○	○	○
Revealing a friend's secret to someone else.	○	○	○	○	○
Leaving your young children alone at home while running an errand.	○	○	○	○	○
Not returning a wallet you found that contains 200€.	○	○	○	○	○

Please answer the following questions.

What is your age?

What is your gender?

- ○ Male
- ○ Female
- ○ Non-binary

What is your current study major?

In general, how willing are you to take risks?

Not at all willing to take risks ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ Very willing to take risks

Is there anything else you would like to tell us? (optional)

A.4 Additional Tables, Analysis of Control Variables, and Manipulation Checks

Table 30: Pairwise treatment comparisons for cheating frequency and control variables

	H vs. M	**H vs. HB**	**M vs. MB**	**HB vs. MB**
Cheating frequency	0.01	0.65	0.66	2.72
	(0.935)	(0.421)	(0.417)	(0.099)
Female	0.17	0.14	0.28	0.25
	(0.679)	(0.703)	(0.598)	(0.619)
Field of Study	2.63	0.12	5.59	5.42
	(0.269)	(0.942)	(0.061)	(0.067)
Risk	-0.92	0.17	0.61	0.47
	(0.359)	(0.871)	(0.548)	(0.642)
Ethical sensitivity	-0.26	-1.04	0.47	1.29
	(0.797)	(0.300)	(0.643)	(0.200)
Closeness	-1.89	-0.40	-0.01	1.22
	(0.060)	(0.6905)	(1.000)	(0.223)
Affinity to technology interaction	1.01	-0.42	-2.15	1.09
	(0.317)	(0.679)	(0.031)	(0.277)
Verification by preferred entity	0.67	0.40	0.41	0.21
	(0.414)	(0.528)	(0.523)	(0.645)
Verification by more error-prone entity	47.23	0.03	0.71	52.06
	(0.000)	(0.859)	(0.401)	(0.000)
Verification by higher discretion entity	77.35	1.29	0.01	62.23
	(0.000)	(0.256)	(0.938)	(0.000)

Pairwise treatment comparisons of control variables. χ^2-values from Pearson χ^2-tests reported for variables female, field of study, verification by preferred entity, verification by more error-prone entity and verification by higher discretion entity with p-values reported in parenthesis. $|z|$-values from Two-sided Mann-Whitney U-tests for variables age, risk, ethical sensitive, closeness and affinity to technology interaction with p-values reported in parenthesis.

Table 31: Comparisons of control variables, by reporting behavior

	Reporting behavior					
	Cheaters			Honest	Comparison	
	Average/ Fraction	Relation to magnitude		Average/ Fraction		
		ρ or χ^2	p		$\lvert z\rvert$ or χ^2	p
Age	22.3 (3.6)	0.156	0.163	21.7 (3.4)	1.04	0.298
Risk	6.4 (2.1)	0.355	0.001	5.5 (2.2)	2.62	0.009
Ethical sensitivity	4.2 (0.4)	0.121	0.279	4.2 (0.5)	0.15	0.881
Closeness	2.9 (1.4)	0.075	0.502	2.8 (1.3)	0.69	0.492
Affinity to technology interaction	3.8 (1.0)	0.028	0.861	3.5 (0.9)	2.11	0.034
Male	0.586	0.67	0.508	0.307	13.35	0.000
Verification by preferred entity	0.634	1.48	0.142	0.580	0.53	0.467
Verification by more error-prone entity	0.488	−0.22	0.833	0.602	2.24	0.134
Verification by higher discretion entity	0.512	1.93	0.053	0.557	0.34	0.560

Standard deviations reported in parenthesis. Comparisons using Pearson χ^2-tests for nominally scales variables (gender, entity-perception variables), Mann-Whitney U-tests for ordinally scaled variables. Spearman's ρ reported for the variables' relation to overreporting magnitude among cheaters. Cheaters: n = 88, Non-cheaters: n = 82.

Table 32: Verification entity preference, perceived error-proneness and perceived decision discretion by treatment

	Human	**Machine**	**Human Black Box**	**Machine Black Box**	**χ^2-test**
Higher perceived error-proneness					
Human	43	34	39	34	$p = 0.681$
Machine	5	7	4	4	
Binomial test	0.000	0.000	0.000	0.000	
Higher perceived decision discretion					
Human	47	39	40	36	$p = 0.739$
Machine	1	2	3	2	
Binomial test	0.000	0.000	0.000	0.000	
Preference for verification entity					
Human	31	18	25	14	$p = 0.041$
Machine	17	23	18	24	
Binomial test	0.059	0.533	0.360	0.143	

Summary statistics of subjects' preferences for verification entity, entities' perceived error-proneness and entities' perceived decision discretion by treatment in absolute frequencies. p-values of Binomial tests for 50/50 response distribution - that would indicate indifference - reported by group per variable. p-values of chi-squared test for distribution between groups reported by variable.

Table 33: Effect sizes (Cohen's d) and post-hoc power tests for pairwise group comparisons

Comparison groups		Frequency		Magnitude	
		d	$1-\beta$	d	$1-\beta$
Human	Machine	0.017	0.035	-0.120	0.069
Human	Human Black Box	0.168	0.096	0.673	0.577
Machine	Machine Black Box	0.181	0.110	1.503	0.999
Human Black Box	Machine Black Box	-0.369	0.367	-0.751	0.662

Group sizes: H: n = 48; M: n = 41; HM: n = 43; MB: n = 38.
Instances of dishonest reporting: H: n = 23; M: n = 20; HM: n = 17; MB: n = 22.
Cohen's d calculated with bootstrapped standard errors for effect sizes.

Table 34: OLS Regression for Likelihood of Cheating (Linear Probability Model)

	Dependent variable: Likelihood of cheating				
	(1)	(2)	(3)	(4)	(5)
Intercept	0.479***	0.419	-0.079	0.322	-0.441
	(0.073)	(0.265)	(0.419)	(0.217)	(0.527)
Treatment					
Machine	0.009	0.026	-0.082	0.147	0.030
	(0.108)	(0.101)	(0.311)	(0.201)	(0.366)
Human Black Box	-0.084	-0.075	-0.089	-0.069	-0.074
	(0.105)	(0.099)	(0.104)	(0.106)	(0.098)
Machine Black Box	0.100	0.105	-0.021	0.228	0.110
	(0.110)	(0.112)	(0.341)	(0.203)	(0.397)
Age		0.010			0.006
		(0.012)			(0.011)
Female		−0.283***			−0.307***
		(0.076)			(0.090)
Field of Study					
Cultural & social studies		0.012			0.032
		(0.086)			(0.091)
Natural science		-0.110			-0.146
		(0.129)			(0.121)
Risk			0.041*		0.042*
			(0.018)		(0.017)
Ethical sensitivity			0.024		0.140
			(0.079)		(0.081)
Closeness			0.007		-0.002
			(0.028)		(0.026)
Verification by machine # ATI					
0			0.057		-0.015
			(0.061)		(0.068)
1			0.080		0.021
			(0.055)		(0.065)
Verification by preferred entity				0.044	0.089
				(0.081)	(0.078)
Verification by more error-prone entity				-0.099	-0.053
				(0.125)	(0.123)
Verification by higher discretion entity				0.222	0.191
				(0.158)	(0.165)
F-test	0.92	2.69*	2.14*	1.07	3.07***
R^2	0.0161	0.1008	0.0708	0.0304	0.1558
Adj. R^2	-0.0017	0.0620	0.0246	-0.0052	0.0736
N	170	170	170	170	170

Note: Coefficients estimated using robust standard errors, standard errors in parentheses; $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.
Model specifications: (1) treatment variables only, (2) including demographics, (3) including control variables, (4) including entity perceptions, (5) full model.

Table 35: Logit regression for likelihood of cheating - Coefficients

	Dependent variable: Likelihood of cheating				
	(1)	(2)	(3)	(4)	(5)
Intercept	-0.083	-0.441	-2.480	-0.825	-4.414
	(0.290)	(1.196)	(1.836)	(1.012)	(2.541)
Treatment					
Machine	0.035	0.121	-0.352	0.697	0.188
	(0.427)	(0.434)	(1.342)	(0.951)	(1.717)
Human Black Box	-0.342	-0.340	-0.383	-0.287	-0.332
	(0.426)	(0.433)	(0.436)	(0.431)	(0.448)
Machine Black Box	0.402	0.468	-0.087	1.030	0.550
	(0.439)	(0.493)	(1.451)	(0.970)	(1.850)
Age		0.047			0.027
		(0.054)			(0.054)
Female		-1.199			-1.406
		(0.336)			(0.439)
Field of Study					
Cultural & social studies		0.056			0.176
		(0.372)			(0.410)
Natural science		-0.510			-0.720
		(0.592)			(0.122)
Risk			0.174		0.195
			(0.077)		(0.081)
Ethical sensitivity			0.106		0.654
			(0.345)		(0.385)
Closeness			0.029		-0.017
			(0.116)		(0.123)
Verification by machine # ATI					
0			0.242		-0.093
			(0.263)		(0.314)
1			0.342		0.087
			(0.240)		(0.293)
Verification by preferred entity				0.180	0.421
				(0.325)	(0.357)
Verification by more error-prone entity				-0.412	-0.256
				(0.510)	(0.600)
Verification by higher discretion entity				1.015	0.975
				(0.839)	(0.914)
Wald $\chi^2(3)$	2.68	14.89	13.63	5.07	28.18
Pseudo R^2	0.012	0.075	0.053	0.535	0.120
N	170	170	170	170	170

Note: Coefficients estimated using robust standard errors, standard errors in parentheses. Model specifications: (1) treatment variables only, (2) including demographics, (3) including control variables, (4) including entity perceptions, (5) full model.

Table 36: Logit regression for likelihood of cheating - Marginal effects

	Dependent variable: Likelihood of cheating				
	(1)	(2)	(3)	(4)	(5)
Treatment					
Machine	0.009	0.027	-0.082	0.166	0.040
	(0.935)	(0.779)	(0.789)	(0.434)	(0.912)
Human Black Box	-0.089	-0.076	-0.089	-0.065	-0.070
	(0.420)	(0.431)	(0.377)	(0.509)	(0.460)
Machine Black Box	0.100	0.106	-0.020	0.242	0.117
	(0.356)	(0.335)	(0.952)	(0.237)	(0.764)
Age		0.010			0.006
		(0.383)			(0.665)
Female		-0.269^{***}			-0.296^{***}
		(0.000)			(0.000)
Field of Study					
Cultural & social studies		0.012			0.037
		(0.881)			(0.665)
Natural science		-0.112			-0.144
		(0.371)			(0.185)
Risk			0.041^{*}		0.041^{*}
			(0.017)		(0.013)
Ethical sensitivity			0.025		0.138
			(0.758)		(0.079)
Closeness			0.007		-0.004
			(0.805)		(0.889)
Verification by machine			0.085		0.137
			(0.299)		(0.632)
ATI			0.067		-0.002
			(0.100)		(0.971)
Verification by preferred entity				0.044	0.089
				(0.578)	(0.233)
Verification by more error-prone entity				-0.100	-0.054
				(0.412)	(0.646)
Verification by higher discretion entity				0.246	0.205
				(0.219)	(0.282)

Note: p-values in parentheses; $^{*}\,p < 0.05$; $^{**}\,p < 0.01$; $^{***}\,p < 0.001$.
Model specifications: (1) treatment variables only, (2) including demographics, (3) including control variables, (4) including entity perceptions, (5) full model.

A.5 Verification Process Illustrations and Outcomes

Human verification

The pictograms displayed in Figure 21 were included in the experimental instructions to illustrate the Verification Part conducted by a human.

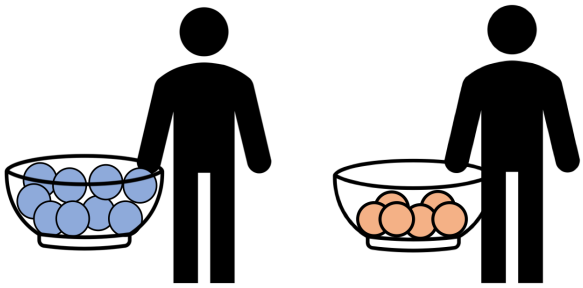


Figure 21: Illustration of human verification process

Algorithmic verification

Translation of text in Figure 22:

"Decision whether your drawn card is checked"

"Please enter your reported number:"

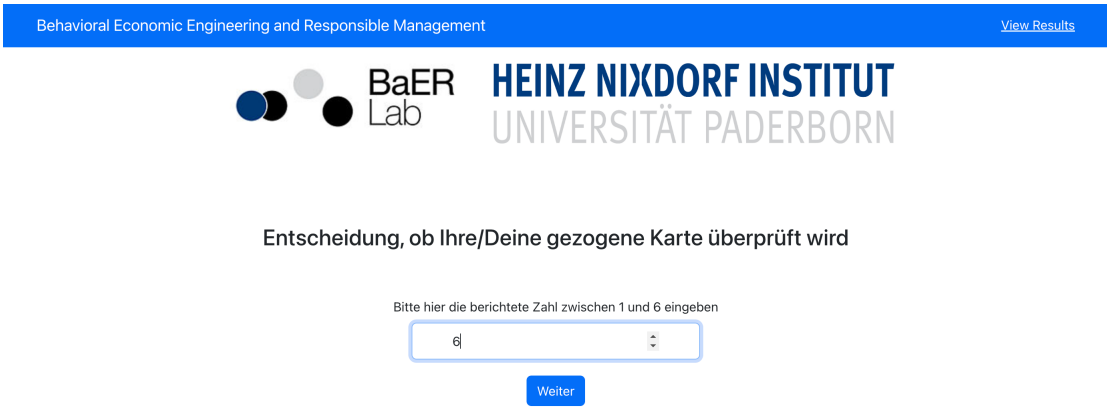


Figure 22: Example for the algorithmic verification interface: Input number 6 (notation in German)

Translation of text in Figure 23:

"Decision whether your drawn card is checked"

"Reported number: 6:"

"Number drawn by the computer: 5"

"The computer's number is lower than or equal to your reported number. Therefore, your drawn card will be checked."

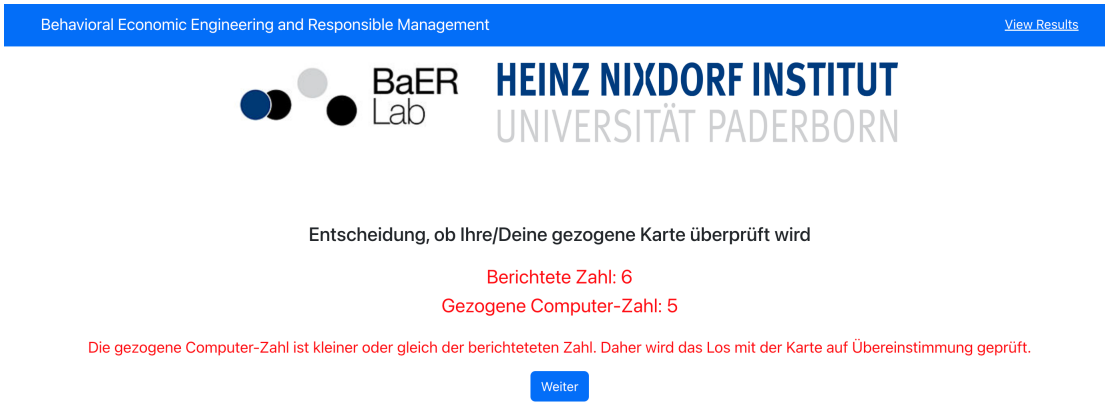


Figure 23: Example for the algorithmic verification interface: Decision on inspection of drawn card (notation in German)

Translation of text in Figure 24:

"Check"

"Reported number: 6"

"Tolerance number: 3"

"If the number on your card is one of the following, you receive the payoff according to your reported number: 3, 4, 5, 6"

"If the number on your card is one of the following, your payoff is reduced: 1, 2"

"Please show your card to the experimenter now."

Figure 24: Example for the algorithmic verification interface: Decision on payoff reduction (notation in German)

Audit outcomes by session

Table 37 displays event sequences and outcomes of the Verification Part for each experiment session.

Table 37: Verification process, by session

Session	Reported	Lottery 1	Card checked	Lottery 2	Reduction	Final payoff
1	2	3	No	-	-	€30
2	4	6	No	-	-	€60
3	2	7	No	-	-	€30
4	6	4	Yes	1	No	€90
5	2	5	No	-	-	€30
6	2	1	No	-	-	€30
7	5	7	Yes	4	Yes	€15
8	2	5	No	-	-	€30
9	6	4	Yes	4	Yes	€15
10	2	8	No	-	-	€30

Supplementary Materials to Chapter 4

B.1 Experiment Instructions

[Translated from German]

Scenario

- Over the course of the experiment, you assume the role of a skilled worker in an industrial company. You will be responsible for the operation of a production facility.
- There is a certain probability that a malfunction may occur in the production facility. To avoid potential malfunctions, maintenance can be performed.
- Your task is to evaluate the probability of malfunctions in multiple rounds and then decide whether the production plant should be maintained in a given round. You will be supported in your task by an artificial intelligence (AI).

Malfunction probability

- The probability of a malfunction is unknown but can be estimated using three indicators.
- These indicators are: Temperature, Speed and Voltage. Each of the indicators can take values between 0 and 100.
- Each of the indicators has its own optimal range. If an indicator is in its optimal range, this is particularly good for the production plant corresponds to a malfunction being less likely.
- The more indicators' values are located outside their respective optimum ranges, the more likely a malfunction becomes.
  - If all three indicators are within their optimal ranges and none are outside, a malfunction is very UNlikely.
  - If two indicators are within their optimal ranges and one is outside, a malfunction is UNlikely.
  - If one indicator is within their optimal range and two are outside, a malfunction is likely.
  - If all three indicators are outside their optimal ranges and none inside, a malfunction is very likely.

Example: The graphic above displays an example for the optimal ranges (green bars) for the three indicators (orange dots). In this example, the "Temperature" indicator is located outside its optimal range and the "Speed" and "Voltage" indicators are located inside their respective optimal ranges. Accordingly, a malfunction would be considered **unlikely** in this case.

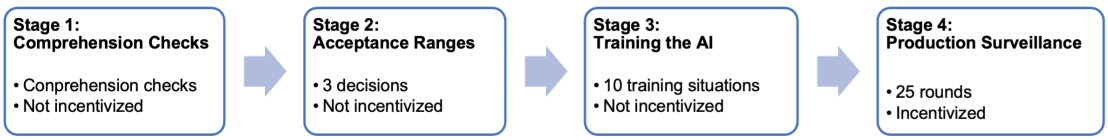
- Important: In the experiment, you do not know the optimal ranges. Instead, you must estimate them as accurately as possible, based on the data points of past malfunctions. This estimation is called "acceptable ranges" (see "Procedure").

Support by an AI

- In each round, the AI predicts the probability of a malfunction and, based on this prediction, gives you a non-binding recommendation as to whether maintenance should be performed.
- The accuracy of the AI predictions can vary. It depends on how the AI has been trained. Training the AI is part of the experiment (see below). You will be informed about the achieved accuracy of the AI (in percent) in the experiment at the end of Stage 3.

Procedure

The experiment consists of four stages that build upon each other.

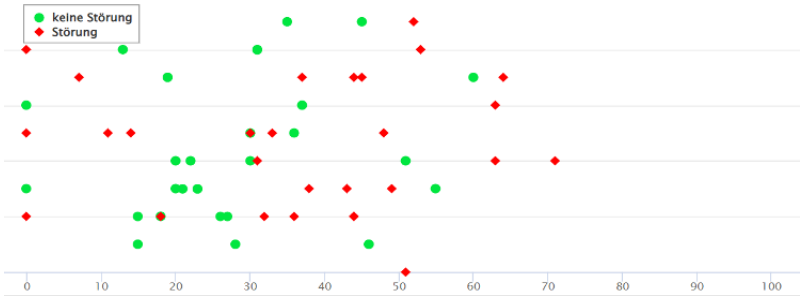


Stage 1: Comprehension checks

- In this stage, comprehension checks are conducted about the instructions. Only once you have answered all the control questions correctly the experiment can begin. You have an unlimited number of attempts to answer the questions correctly.

Stage 2: Selection of acceptable ranges

- As mentioned, the optimal ranges of the individual indicators are unknown to you. Instead, you must define an acceptable range for each indicator.
- An acceptable range is an approximation of the actual (unknown) optimal range. The closer the acceptable ranges you set are to the optimal range, the better the AI's advice will be.
- Data about past malfunctions is available for you to set your acceptable ranges:
  - For each indicator individually, you can see at which values there were malfunctions in the past (red dots) and at which there were not (green dots).
  - You are now asked to define a lower limit (minimum) and an upper limit (maximum) of your acceptable range (green dashes). In general, an acceptable range should contain as many points without malfunctions (green) and as few points with malfunctions (red) as possible.

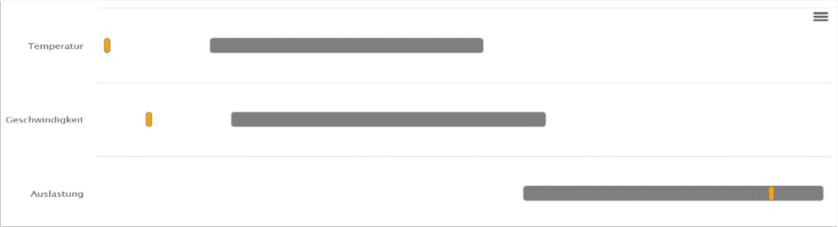


  - At the beginning, you will be given an example that you can use to practice setting the limits (technical note: the limit that is closer to your mouse pointer moves in each case).
  - After you have set and confirmed the acceptable ranges, they will be displayed as gray bars in the further course of the experiment for the sake of conciseness (see figure). The individual data points are hidden.

- You define a total of three acceptable ranges (one for each indicator), which you will necessarily need in the further course of the experiment.
- Your acceptable ranges will be displayed for all further decisions, so you do not have memorize or note them.

Stage 3: Training the AI

- In this stage, the AI is trained based on your acceptable ranges defined in Stage 2. The AI thus learns how to evaluate the probability of malfunctions for different indicator combinations.
- The training of the AI happens through ten training situations as follows:
  - Each training situation represents a combination of the three indicators' values. These values are shown together with their acceptable ranges.
  - The following figure provides an example of a training situation. Orange bars represent the indicators' values. Gray bars represent the acceptable ranges.



  - For each training situation, you can see which indicators are within and which are outside your defined acceptable ranges.
  - Your task is to tell the AI how each training situation is to be evaluated regarding the likelihood of a malfunction. In doing so, you help the AI learn.
  - Use your acceptable ranges and your knowledge about the probability of malfunctions for the evaluation:
    - If three indicators are within your acceptable ranges and zero are outside, a malfunction is very UNlikely.
    - If two indicators are within your acceptable ranges and one is outside, a malfunction is UNlikely.
    - If one indicator is within your acceptable ranges and two are outside, a malfunction is likely.
    - If there are zero indicators inside your acceptable ranges and three outside, a malfunction is very likely.
  - Each of your ten malfunction probability assessments is then checked for correctness:
    - If a malfunction has been classified as "very unlikely" or "unlikely" and no malfunction has actually occurred, the assessment is considered correct and otherwise incorrect.

    - If a malfunction was classified as "very likely" or "likely" and a malfunction actually occurred, the evaluation is considered correct and otherwise incorrect.
    - Whether a malfunction actually occurs or not depends on the actual optimum ranges, which remain unknown.

- The result of the training, and thus the quality of the AI, depends on how many training situations have been correctly assessed. You will be informed about the result at the end of the training stage. Two results are possible:
    - If at least seven training situations were evaluated correctly, you will receive an AI with the accuracy of 90% (on average it is correct in 9 out of 10 cases and wrong in one out of 10 cases).
    - If less than seven training situations were evaluated correctly, you will receive an AI with the accuracy of 50% (it is correct on average in 5 out of 10 cases and wrong in 5 out of 10 cases).
- After completing this stage, the AI has learned to evaluate malfunction probabilities in comparable situations through the training situations.
- In stage 4, you can use the AI for decision support.

Stage 4: Production plant surveillance

- This stage consists of 25 rounds.
- In each round, you have to make the decision whether to maintain the production facility.
- All rounds are independent of each other, i.e., the decision in one round does not affect other rounds.
- In each round, you will receive a graphic showing the values of the three indicators and your self-defined acceptable ranges (see Stage 3).
- In In each round you make your decision in two steps:
    - In the first step, you evaluate the given situation in terms of the probability of failure and decide whether maintenance should be performed.
    - In the second step, the AI's recommendation is displayed to you. Afterwards, you are asked again whether you want to perform maintenance.
- Only the decision in the second step is relevant for your payoff in the respective round.

- Whether a malfunction actually occurs or not depends on the optimal ranges, which remain unknown. You will only find out at the end of the experiment how often you were correct and how high your payoff will be.

Payoffs

- During the experiment, all amounts are denoted in the fictitious currency "Taler".
- Per round, depending on your maintenance decision and the occurrence/non-occurrence of a malfunction, you will receive the following payoffs:
  - You decide that maintenance should be performed.
    - Maintenance limits your production capacities. Therefore, your payoff this round is 5 Taler.
  - You decide that no maintenance should be performed.
    - If no malfunction occurs and you can therefore produce fully, your payoff from this round is 10 Taler.
    - If a malfunction occurs and therefore you cannot produce, your payoff from this round is 0 Taler.
- The payoffs from all rounds are cumulated.
- At the end of the experiment, you will receive your payoffs at an exchange rate of 1 € per 10 Taler. In addition, you will receive a show-up fee of 2.50 €.

Additional remarks

- All communication is prohibited for the duration of the experiment except for communication explicitly permitted by the instructions.
- Mobile phones must be turned off for the duration of the experiment.
- All decisions within the scope of the experiment will remain completely anonymous.
- After completing the main part of the experiment, we kindly ask you to answer some additional questions. Answering the questions honestly and in full is very important for the subsequent analysis of the experiment. The answers to the questions remain anonymous and will only be evaluated for scientific purposes. Your answers in this questionnaire have no impact on your payoff achieved in the experiment.

B.2 Questionnaire

Please answer the following questions.

What is your age?

What is your gender?

- Male
- Female
- Non-Binary

What is your highest level of education?

- Highschool / GED
- Undergraduate degree
- Graduate degree
- Else / Prefer not to say

What is your current study major?

Please answer the following questions.

Please indicate your consent with the following statement on a scale from 1 (= completely disagree) to 7 (= completely agree).

	1	2	3	4	5	6	7
I believe that my estimates of malfunction probabilities were correct.	○	○	○	○	○	○	○
I believe that my estimates of malfunction probabilities are close to the true value.	○	○	○	○	○	○	○
I was very confident about the accuracy of my estimates of malfunction probabilities.	○	○	○	○	○	○	○
I was confident I would do well on the task.	○	○	○	○	○	○	○
I have no doubt that my estimates of malfunction probabilities are close to the true values.	○	○	○	○	○	○	○

Please answer the following questions.

Please indicate to what extent the following statements apply to you personally.

	Does not apply at all	Applies little	Somewhat applies	Pretty much applies	Fully applies
In difficult situations I can rely on my abilities.	○	○	○	○	○
I can cope well with most of the problems on my own power.	○	○	○	○	○
Even strenuous and complicated tasks I can usually solve well.	○	○	○	○	○

Please answer the following questions.

Please indicate your consent with the following statement on a scale from 1 (= Do not consent at all) to 5 (= Fully consent).

	1	2	3	4	5
I could easily understand the maintenance scenario presented in the experiment.	○	○	○	○	○
I was able to put myself in the role of a production plant manager.	○	○	○	○	○
The information and examples provided to me prepared me well for the decisions in the experiment.	○	○	○	○	○
I could understand most of the explanations in the experiment.	○	○	○	○	○
I understood the contexts presented in the experiment.	○	○	○	○	○
I knew exactly what was required of me in the experiment.	○	○	○	○	○
I knew what I had to do to be as successful as possible in the experiment.	○	○	○	○	○
The tasks of the experiment were demanding.	○	○	○	○	○
I am satisfied with my performance in the experiment.	○	○	○	○	○
I think most of the experiment participants did well on the tasks.	○	○	○	○	○

Please answer the following questions.

Please indicate your consent with the following statement on a scale from 1 (= Do not consent at all) to 5 (= Fully consent).

	1	2	3	4	5
The AI-training stage was time well spent.	○	○	○	○	○
I would have liked to invest more time in training the AI.	○	○	○	○	○
I was instrumental in how good the quality of the AI recommendations was.	○	○	○	○	○
I felt like it was a gamble whether you got a high-quality AI or a low-quality AI at the end of the training stage.	○	○	○	○	○
Training the AI has helped me better understand how the AI makes its recommendations.	○	○	○	○	○
Training the AI helped me better understand how it works.	○	○	○	○	○
I liked that my expert knowledge was shared with the AI in the form of my acceptable intervals.	○	○	○	○	○
I found it uncomfortable that the AI had access to my expert knowledge in the form of my acceptable intervals.	○	○	○	○	○
If I had the chance, I would have preferred to share even more knowledge with the AI.	○	○	○	○	○
If I had a choice, I would rather not have shared any knowledge with the AI at all.	○	○	○	○	○
My influence on training the AI was great.	○	○	○	○	○
I enjoyed training the AI.	○	○	○	○	○
I found the AI recommendations to be of high quality.	○	○	○	○	○
When deciding whether to perform maintenance, I weighed things logically.	○	○	○	○	○
I made decisions about whether to perform maintenance intuitively rather than strategically.	○	○	○	○	○

Please answer the following questions.

In the following questionnaire, we will ask you about your interaction with technical systems. The term "technical systems" refers to apps and other software applications, as well as entire digital devices (e.g., mobile phone, computer, TV, car navigation).

Please indicate to what extent you agree to the following statements.

	Completely disagree	Largely disagree	Slightly disagree	Slightly agree	Largely agree	Completely agree
I like to occupy myself in greater detail with technical systems.	○	○	○	○	○	○
I like testing the functions of new technical systems.	○	○	○	○	○	○
I predominantly deal with technical systems because I have to.	○	○	○	○	○	○
When I have a new technical system in front of me, I try it out intensively.	○	○	○	○	○	○
I enjoy spending time becoming acquainted with a new technical system.	○	○	○	○	○	○
It is enough for me that a technical system works; I don't care how or why.	○	○	○	○	○	○
I try to understand how a technical system exactly works.	○	○	○	○	○	○
It is enough for me to know the basic functions of a technical system.	○	○	○	○	○	○
I try to make full use of the capabilities of a technical system.	○	○	○	○	○	○

B.3 Additional Tables and Figures

Table 38: Illustration of Decision Process for Each Round of Maintenance Stage

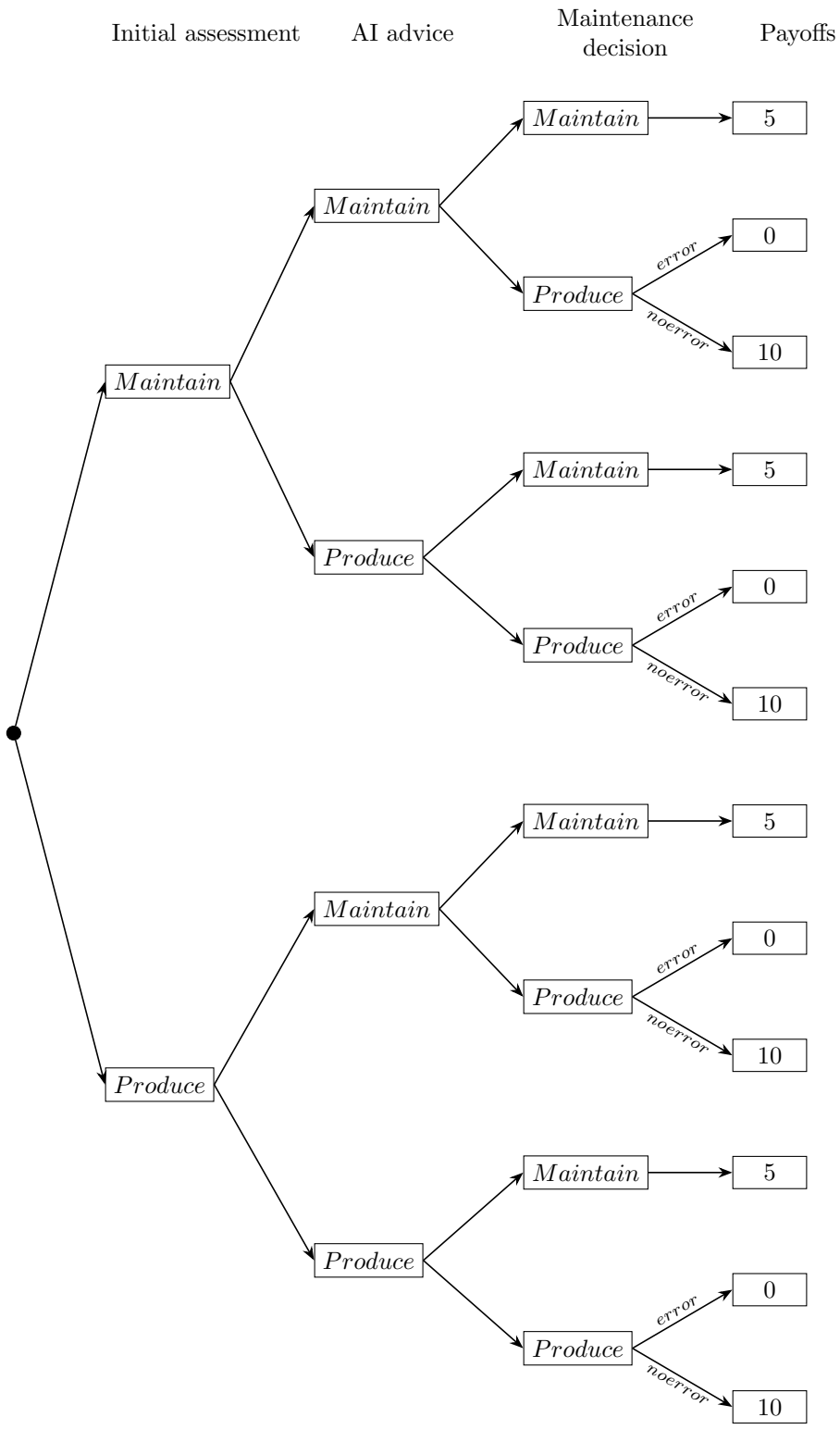


Table 39: Maintenance Recommendation by the AI in each round of the Maintenance Stage (high-accuracy AI)

Round	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Maintenance recommendation	no	yes	yes	yes	yes	no	no	no	yes	no	yes	no	no	no	no	no	no	no	no	no	yes	no	yes	no	no
Resulting payoff	10	5	5	5	5	0	10	10	5	0	5	10	10	10	0	10	10	10	0	10	5	10	5	10	10

Note: This table reports the (high-accuracy) AI's maintenance recommendations for each round of the maintenance stage, as well as the corresponding payoffs. As the high-accuracy AI is 90% accurate, errors occur in a total of four round, i.e., no maintenance is recommended despite being necessary, therefore the respective round's payoff is zero.

Table 40: Distribution of Study Majors Among Participants

	Frequency	Percentage
Study Major		
Arts & Design	13	8.50
Business & Economics	52	33.99
Computer Science	11	7.19
Cultural & Social Studies	11	7.19
Engineering	10	6.54
Natural Sciences	1	0.65
Pedagogics	53	34.64
Non-student	2	1.31
Total	153	100.00

Table 41: Efficient Revision Rate and AI Net Benefit

	$\mathbf{T}_Z$	$\mathbf{T}_P$	$\mathbf{T}_A$	K-Wallis
Efficient revision rate				
Mean	0.73	0.82	0.70	0.231
Standard deviation	0.26	0.24	0.34	
Min	0	0.33	0	
Median	0.73	1	0.77	
Max	1	1	1	
Number of observations	40	37	35	
AI net benefit				
Mean	0.86	0.87	0.78	0.487
Standard deviation	0.19	0.21	0.29	
Min	0.5	0.33	0	
Median	1	1	1	
Max	1	1	1	
Number of observations	37	36	33	

Note: In total, 30 subjects did not revise their initial decision intent at least once. Therefore, they were excluded from this calculation, as could not be computed for them. Analogously, a total of 36 subjects either did not revise their initial decision intent at least once or follow the AI's advice at least once. Therefore, they were excluded from this calculation, as an AI net benefit could not be computed for them. "K-Wallis" reports the p-value for Kruskal-Wallis H-Tests with ties between experimental groups.

Table 42: Summary Statistics and Between-Group Comparison of Treatment Perception

	$\mathbf{T}_Z$	$\mathbf{T}_P$	$\mathbf{T}_A$	K-Wallis
Items on treatment perception				
AI-training was time well spent	4.02	4.02	3.72	0.5221
	(0.94)	(0.94)	(1.23)	
Would have liked to spend more time for training	3.22	2.73	2.77	0.1388
	(1.27)	(1.34)	(1.52)	
Perceived contribution to AI-advice quality	3.34	3.87	4.11	0.0013
	(1.10)	(1.06)	(0.91)	
AI-quality outcome perceived as gamble	2.52	2.13	1.96	0.1163
	(1.31)	(1.12)	(0.91)	
Perceived understanding of AI's functionality	3.08	3.38	3.91	0.0024
	(1.21)	(1.23)	(1.02)	
Perceived understanding of how AI generates advice	2.90	3.42	3.72	0.0020
	(1.15)	(1.16)	(1.16)	
Liked sharing knowledge with AI	4.00	3.89	3.96	0.8598
	(0.78)	(0.91)	(0.88)	
Felt uncomfortable sharing knowledge with AI	1.54	1.51	1.40	0.8251
	(0.99)	(0.84)	(0.83)	
Would have liked to share more knowledge with AI	3.52	3.42	3.34	0.7511
	(1.36)	(1.16)	(1.37)	
Would have preferred not to share any knowledge with AI	1.54	1.64	1.64	0.9923
	(0.76)	(0.98)	(1.01)	
Perceived influence on AI-training	3.02	3.58	4.09	0.0001
	(0.91)	(1.12)	(0.80)	
Enjoyed training the AI	3.26	3.42	3.70	0.1646
	(1.19)	(1.03)	(1.08)	
Perceived quality of AI-advice	3.68	3.76	3.60	0.4874
	(0.79)	(1.05)	(0.95)	
Maintenance decisions made rather strategically	4.28	4.44	4.26	0.3548
	(0.76)	(0.62)	(0.64)	
Maintenance decisions made rather intuitively	2.28	2.09	1.98	0.3967
	(1.09)	(1.04)	(0.92)	
Number of observations	50	45	47	

Note: This table reports summary statistics of questionnaire items on treatment perception (5-point scale) by treatment. Standard deviations reported in parenthesis. "K-Wallis" reports p-values for Kruskal-Wallis H-Tests with ties between experimental groups. Only subjects with high-accuracy AI included.

Table 43: Pairwise between-group comparisons of standardized questionnaire scores

	$\mathbf{T}_Z$ vs. $\mathbf{T}_P$	$\mathbf{T}_Z$ vs. $\mathbf{T}_A$	$\mathbf{T}_P$ vs. $\mathbf{T}_A$
Affinity to Technology Interaction (ATI)	0.477	2.514	1.618
	(0.6359)	(0.0115)	(0.1063)
Self-efficacy (ASKU)	0.160	1.167	1.532
	(0.8749)	(0.2452)	(0.1264)
Ex-post decision confidence (DC)	0.606	0.167	0.646
	(0.5476)	(0.8698)	(0.5211)

Note: This table reports results for pairwise two-sample between-group Mann-Whitney U-Tests between experimental groups. $|z|$-values reported with p-values in parenthesis.

Supplementary Materials to Chapter 5

C.1 Experimental Research Method and Induced Value Theory

The deviation between theoretically prescribed behavior and actual human decision making is often labeled "behavioral messiness" and poses a huge challenge for organizational policy makers. Experiments thereby represent a tool for bridging economic theory and real-world institutional design in order to generate practical value from applied economic science [Bolton and Ockenfels, 2012].

Behavioral economic findings are often based on experimental research, especially controlled laboratory experiments, as "experimental control is exceptionally helpful for distinguishing behavioral explanations from standard ones" [Camerer and Loewenstein, 2003]. Laboratory experiments, which had been considered unfeasible for economic disciplines and privilege of natural sciences up until the late 20th century [Samuelson and Nordhaus, 1985], aim to parallel real-world situations in laboratory settings, while abstracting from environmental factors. This means, experiments isolate certain variables of interest from more complex real-world contexts, while simultaneously controlling for conditions of the subjects' economic and social environments [Roth, 1988]. Thereby, the laboratory acts as a "wind tunnel" for practice in order to test behavioral reactions to institutional interventions in a controlled environment, before implementing them into the real world. Projects following this approach are summed under the label of "behavioral economic engineering" [Bolton and Ockenfels, 2012]. Prominent examples include the implementation of optimized retirement savings plans [Thaler and Benartzi, 2004] and the redesign of matching algorithms for American physicians Roth and Peranson [1999].

Participants, commonly referred to as "subjects", work on computerized or analogue tasks, like solving math problems [Mazar et al., 2008] or assembling Lego figures [Ariely et al., 2008], accompanied by anonymously making decisions, which are observed by the experimenter. Economic theories or concepts that the experimenter wants to test usually underlie the decision to make in the experiment often in relation to a given task. A constituting factor of experiments consists in them being incentivized, i.e., the participants are paid for solving tasks and making decisions by the experimenter, with the concrete amount of payments depending on the respective experimental design [Roth, 1988]. Subjects know about the payments they are able to receive beforehand, as they receive instructions containing the experimental procedure and the payoff function in the beginning of every experiment. Herein lies the biggest advantage of experimental economic research: actual human behavior can be observed with the subjects' actions having actual monetary consequences for their payoffs in the end, so they have to "put their money where their mouth is". This is not the case for other research methods used in economic or social science like surveys or scenario studies, in which participants only

state how they would behave in certain situations, while no information is gained whether they would actually behave the way they stated when faced with the decision-problem in reality. The problems of intention-behavior gaps [Carrington et al., 2010] and giving socially desirable answers [Nederhof, 1985] are well-known.

As a basic principle of economic experiments, there is no deception by the experimenter. Subjects will be asked to do exactly what is stated as their task in the instructions they receive before the experiment and they will be paid exactly according to the payoff function given in the instructions as well [Weimann and Brosig-Koch, 2019]. Experiments can be used for various purposes like testing economic theories and theoretical equilibria, testing policies and environments, establishing phenomena, stylized facts and new theories or deriving political recommendations [Roth, 1988]. Furthermore, experimental evidence can be replicated and re-evaluated by other experimenters using the same experimental setup and instructions, which are usually published alongside the results [Charness, 2010].

In order to experimentally test certain theories or economic interventions, subjects are randomly assigned into one control group and (at least) one treatment group, in order to avoid (self-)selection biases, with a treatment being the intervention that is going to be tested. Differences between control group and treatment group must thereby be reduced to exactly one factor, namely the so-called treatment variable, i.e., the intervention that the experimenter wants to evaluate [Weimann and Brosig-Koch, 2019]. Herein lies another major advantage of controlled economic experiments as they allow causal inferences. Due to the experimental environment being held constant and environmental factors being controlled for all groups, they cancel out through comparative statics between groups. Therefore, outcome in a treatment group compared to a control group can only be caused by an exogenous change, i.e., by the treatment intervention itself since it represents the only factor that differs between groups [Pearl, 2003; Camerer and Loewenstein, 2003; Weimann and Brosig-Koch, 2019].

As mentioned above, a constituent factor of economic experiments consists in the provision of decision-dependent monetary incentives. Economic experiments underlie the general assumption that any kind of utility an individual experiences can be expressed by an equivalent monetary incentive. Through these incentives, the experimenter is able to induce a certain utility function $U(x)$ onto the subjects, in order to neutralize the subjects' inherent preferences for the duration of the experiment. This "Induced Value Theory" was originally introduced by Nobel Memorial Prize in Economic Sciences laureate Vernon Smith [1976]. According to him, subjects should derive all utility from the monetary incentives provided in the experiment itself, during their participation.

While the internal validity of experiments is usually recognized to be very strong, the external validity of these experiments, i.e., the transferability of their results to the world outside the laboratory, the external validity, is commonly discussed and criticized by opponents of this methodology [Friedman

and Sunder, 1994; Guala, 2005]. For instance, in the context of perceptional and judgmental biases, laboratory settings are accused of changing the environment in which a human makes efficient decisions to an artificial one. However, it can be argued the other way around in that laboratory experiments show people at their best through providing all information necessary and eliminating distractions [Slovic et al., 1978]. If people fall to biases in this isolated, save environment, they will likely do so outside of it as well.

C.2 Experiment Instructions

[Translated from German]

Scenario

:

- You own a flight **drone with a photo function**. You are hired by the city administration of your hometown to take pictures of the traffic volume at several traffic junctions with the help of your drone.
- Once the order is completed, you are able to sell the drone at a fixed price in case it is still intact.
- The drone independently flies a predetermined route in each period and **photographs autonomously**, you only have to make the **decision about the number of rounds to fly**.
- You will be paid by the city administration depending on the information value of the pictures taken.

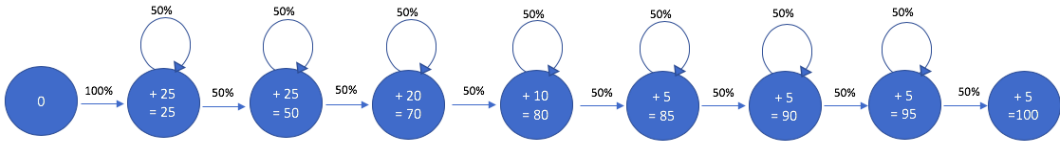
Procedure

- The experiment consists of **10 periods**.
- In each period a new traffic junction is approached. The transfer between the traffic junctions takes place by car.
- For each period a new battery is provided to operate the drone. This allows you to fly a **maximum of 8 laps** over the respective traffic junction, i.e. to take a maximum of 8 pictures.
- The batteries for operating the drone are provided by the city administration, the transfer between the intersections by car is also organized by the city administration. Therefore, you do not have to bear any costs.
- The drone flies and photographs autonomously.
- Your task is to **decide** for each period **how many rounds** the drone should fly.
- The aim of your assignment is to obtain the highest information value possible for each traffic node. In case of a constant improvement in each round, i.e. 8 optimal images, an information value of 100 is obtained.

- In the **first round of each period**, an image with an **information value of 25** is reached **with certainty**.
- From the **second to the eighth round** of each period, the pictures taken improve the information value by a given amount (see table below) with a **constant probability of 50%**.
- With a p**robability of 50%**, however, there is **no improvement** in the information value. This can be caused by uncontrollable external influences such as rain, strong winds, poor lighting conditions or lack of traffic flow.
- You therefore have a 50/50 chance to achieve an increase in information value in each round. If no increase occurs, you remain at the information value you have achieved so far and have the chance to improve the information value to the next higher "level" again in the next round.
- The possible information value gains are listed in the following table:

Round	2	3	4	5	6	7	8
Information gain achieved with 50% probability	25	20	10	5	5	5	5
Maximum information value attainable at this point	50	70	80	85	90	95	100

- The process of improving the information content is summarized in the following graphic:



- In each round of each period (after the picture has been taken), there is also a **constant probability of a drone crash of 2%**.
- For each period, you receive a payment equal to the information value in the fictitious currency Taler (for example, XX Taler for an achieved information value of XX in period Y).
- If the drone is still intact, you are able to sell the drone at a fixed price of 400 Taler in the end of the experiment.
- In case of a crash, you can no longer sell the drone and will not receive 400 Taler.

Feedback on the obtained information value

[Depending on the treatment, the instructions included one of the following paragraphs.]

Open Loop Treatment:

- The drone stores the pictures on a built-in memory chip. In order to retrieve the information value of the images, you have to disassemble the drone after all periods are completed in a complex procedure and read out the memory chip on the computer.
- You only receive **feedback** about the information value achieved in each individual period in the end of the experiment **after all periods are completed**.
- You decide in advance for all periods how many rounds are to be flown in the respective period.

Closed Loop Treatment:

- The drone transmits the pictures automatically and immediately to your notebook after each round flown.
- **After each flown round**, you will receive **feedback** on whether the information value of the pictures has improved during that round and how much information value has been obtained until this point in time.
- After receiving feedback, you decide whether the drone should fly another round or whether the period should end.

End of the experiment

- In the event of a crash, the drone is broken, and the experiment ends immediately. You can't fly in the following periods and you can't sell the drone anymore. However, the pictures taken up to the time of the crash are still usable and will account for your payoffs.
- In case the drone does not crash, you can sell the drone at the end of the experiment for a **fixed price of 400 Taler**.
- The experiment is finished when either all periods have been completed (and the drone is sold) or the drone has crashed.

Questionnaire

- After the experiment is finished, you will be asked to answer some additional questions. The questionnaire is divided into two parts.
- Part 1 of the questionnaire has no impact on your payment.
- Part 2 of the questionnaire offers **every 15th experiment participant** the chance to win an additional prize, which is paid together with the earnings from the experiment.

Payoffs

- Your payoffs consist of the payment for the information value achieved by the drone pictures and the sales revenue of the drone itself.
- In the event of a crash, you will be paid for the information value you have obtained up until this point.
- They are paid at an **exchange rate** of: **1€ per 120 Taler**.
- You will receive your payment at the secretariat of the Chair of Corporate Governance during office hours, by stating your personal participant ID (see below).
- Please collect your payoffs by Thursday, 17 October 2019 at the latest. A later payment cannot be considered.
- You can find the secretariat's opening hours on our homepage.

Participant ID:

- Since you make all decisions in the experiment anonymously, you have to generate a personal ID code, so your generated payoffs can be assigned to you afterwards.
- You will be able to collect your payoffs in the secretariat of the Chair of Corporate Governance (Campus of Paderborn University, Q building, Room Q3.119), by stating your personal participant ID code.
- Your personal participant ID is built as follows:
  - First and last letter of your mother's first name: Example: **A**NN**E** = **AE**
  - First and last letter of your father's first name: Example: **T**HORSTE**N** = **TN**
  - First and last letter of your first name: Example: **M**ICHAE**L** = **ML**
  - Date of birth of your mother: Example: **17** July 1965 = **17**

Good luck and many thanks for participating in this experiment!

Comprehension Checks

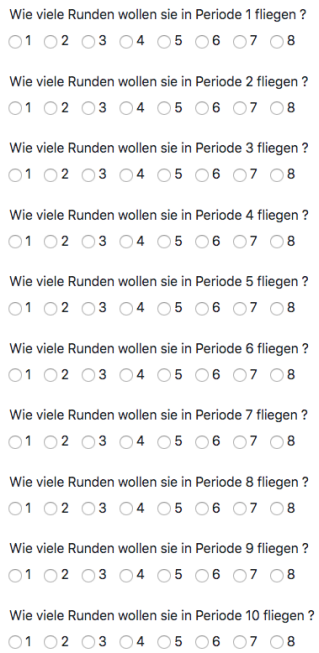
The following control questions were posed to the participants in order to ensure that they had understood the critical aspects of the instructions. Participants were not able to advance to the drone piloting task as long as they did not solve all four questions correctly.

- In the event of a drone crash, you will not be paid for this period: True/False [Answer: No]
- In the first period, an information value of 25 is achieved with certainty. True/False [Answer: True]
- The probability of the drone crashing is constant at ... [Answer: 2]
- The maximum information value to be achieved is ... [Answer: 10]

C.3 Choice and Result Screens

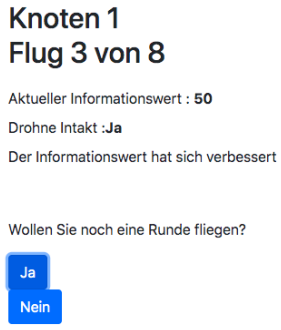
The following information were provided as feedback to the participants in the respective treatments (annotations in German).

Open Loop:



Subjects are asked to choose how many rounds they want to fly in each period. They receive the end results afterwards on a result screen.

Closed Loop:



After each choice, one screen provides immediate feedback on whether the information value has increased in the previous round, on the currently accumulated information value (from this period), and on whether the drone is still intact. Below, the subject can choose whether they want to fly another round ("Ja") or finish the current period and continue with the next one ("Nein").

C.4 Individual Risk Preferences

Individual risk preferences represent a factor commonly influencing human behavior that should necessarily be considered in order to explain individual decision making under risk and uncertainty. Human risk preferences are generally differentiated into risk averse, risk neutral and risk seeking behavior. They are usually determined by the individual's subjective evaluation of a probabilistic payment from a lottery compared to a certainty equivalent, i.e., a safe fix payment [Holt and Laury, 2002; Dohmen et al., 2010]. If an individual subjectively judges the utility from the expected value E of a payment X as higher than the expected utility of such payment, namely $u(X)$, then this individual is classified as risk averse [Pratt, 1964]:

$$u(E[X]) > E[u(X)]$$

As a consequence of Jensen's inequality, a risk averse individual's personal utility function is concavely shaped. This means that the individual is willing to sacrifice the chance of higher but uncertain payments in order to receive a smaller but certain payment, the certainty equivalent, thereby paying the so-called insurance premium [Pratt, 1964]. In an example by Varian [2006] displayed in Figure 25, an individual with €10 at his disposal has to decide whether or not to participate in a lottery in which €5 are gained with a probability of 50% and €5 are lost with a probability of 50%. For a risk averse individual the utility of the expected value $u(10)$ is greater than the expected utility of wealth $0.5u(15) + 0.5u(5)$.

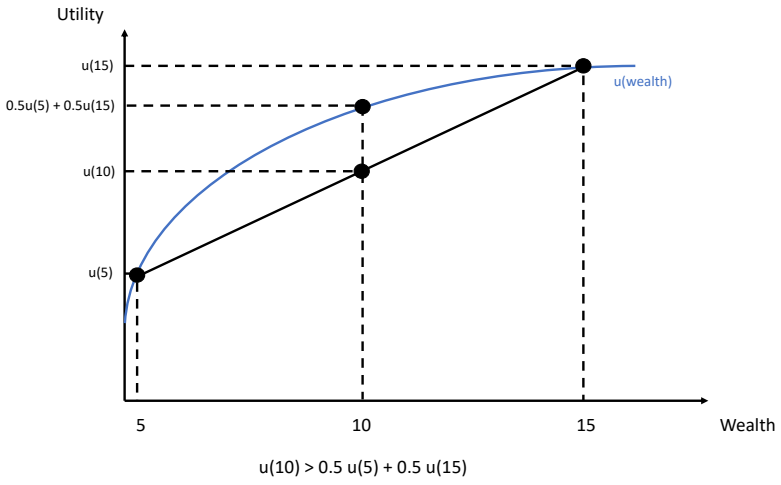


Figure 25: Example for a risk averse individual's utility function, adapted from Varian [2006]

The other way around, if an individual derives less utility from the expected value of X than the X's expected utility

$$u(E[X]) < E[u(X)],$$

then this individual is considered risk seeking. Its subjective utility function is convex (see the analogous example in Figure 26). Risk seeking individuals are willing to give up the certainty equivalent in order to get the chance to play the lottery through which they may obtain a higher payment, i.e., they favor gambling. The payment obtainable beyond the certainty equivalent through playing the lottery is called risk premium. In case an individual is indifferent between both alternatives, the individual can be considered risk neutral, with risk and insurance premium canceling out [Pratt, 1964].

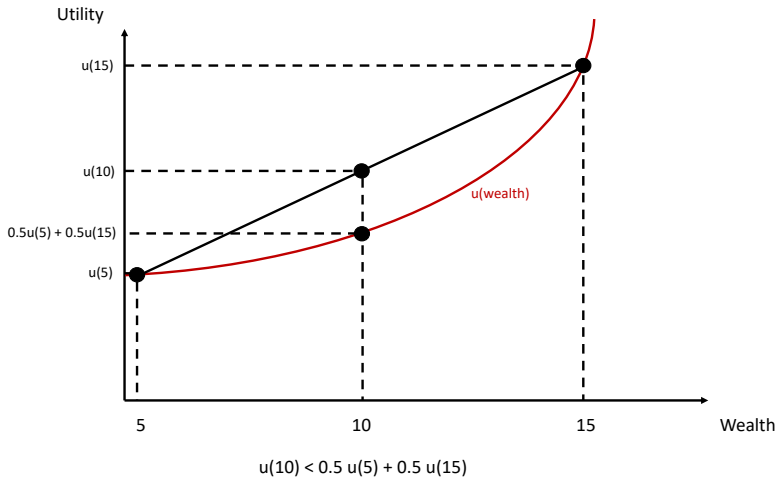


Figure 26: Example for a risk seeking individual's utility function, adapted from Varian [2006]

Standard economic theories, in the tradition of the *homo oeconomicus*, are usually based upon the explicit or implicit assumption of individual risk neutrality, reflecting notion of individuals as rationally optimizing ("Markovian") decision makers [Jaquette, 1976]. Contrarily, behavioral economic research usually observes risk averse decision makers, in the laboratory [Harrison et al., 2007b; Eckel and Grossman, 2008], as well as in the field [Harrison et al., 2007a].

The degree of human risk aversion was found to increase with the size of incentives being at stake [Kachelmeier and Shehata, 1992; Holt and Laury, 2002], contradicting theories of constant relative risk aversion. However, individuals are observed to act clearly risk averse even only with low incentives at stake, wherefore models that assume risk neutral human behavior do not adequately represent reality and may lead to biased inference [Holt and Laury, 2002]. Regardless of this quite clear evidence, only sporadic effort exists in order to incorporate risk preferences other than risk neutrality into Markov Decision Processes [Howard and Matheson, 1972; Majumdar et al., 2017].

Multiple price list formats

Multiple price lists formats offer a simple but effective tool to reveal subjects' actual risk preferences. The basic concept was designed with the goal of testing the standard economic assumption of risk

neutrality as well as the behavioral assumption of constant relative risk aversion. The list consisted of a table with two columns that represented two lotteries the subjects had to choose between for ten rows featuring varying probabilities of winning certain amounts of money that remained constant for all rows [Holt and Laury, 2002]. As the original price list contained monetary values with two decimals and varying probabilities that had to be weighed, it can be assumed that it was difficult for subjects to compare the different options. The subsequent work Dohmen et al. [2010] designed a more clear-cut version of the multiple price list (see Table 44) in which subjects had to choose between a safe payoff that increased by one Euro with each row and a fair lottery that remained constant for all rows.

Table 44: Multiple price list by Dohmen et al. [2010]

	Option A	**Option B**
1)	€0 safe	€30 with a probability of 50%, €0 with a probability of 50%
2)	€1 safe	€30 with a probability of 50%, €0 with a probability of 50%
3)	€2 safe	€30 with a probability of 50%, €0 with a probability of 50%
4)	€3 safe	€30 with a probability of 50%, €0 with a probability of 50%
⋮	⋮	⋮
19)	€18 safe	€30 with a probability of 50%, €0 with a probability of 50%
20)	€19 safe	€30 with a probability of 50%, €0 with a probability of 50%

Subjects are asked to indicate for each row whether they prefer the offered safe payment (Option A) or the offered fair lottery (Option B). The safe payment equals the number of the respective row minus one (in Euro), starting at €0 in the first row and increasing up to €19 in the twentieth row. The lottery is the same in each row, containing a 50% chance to win €30 and a 50% chance of winning nothing. A subject's risk attitude can be determined by observing the row in which his choice changes from B to A. When continuously choosing B until row 16 and continuously choosing A afterwards, a subject is classified as risk neutral, as both options yield the exact same expected payoffs of €15 in row 16. At this point a risk neutral subject is indifferent between the receiving the safe payment and playing the lottery. If the change from B to A occurs before the sixteenth row, then a subject is classified as risk averse, as it is willing to sacrificing the chance to play a lottery that yields a higher expected payoff in favor of a safe payment of a lower amount than such expected payoff. If the change occurs after the sixteenth row, a subject is regarded as risk seeking. Contrarily to risk averse subjects, risk seeking subjects are willing to sacrifice a safe payment in order to be able to play a lottery that yields an expected payoff that is lower than safe payment but yields a higher maximum payoff.

After making their choices for every row, a fraction of the participant group (e.g., every seventh [Dohmen et al., 2010] or fifteenth [Djawadi and Fahr, 2013] subject) is chosen at random and actually paid for one randomly chosen row of the table, according to their respective choice made in this row. This procedure incentivizes subjects to choose in accordance with their true preferences in each row,

since they would forfeit expected utility if they did not [Dohmen et al., 2010]. Therefore, the multiple price list format presents an incentive-compatible instrument to measure individual risk preferences.

Supplementary Materials to Chapter 6

D.1 Discussion on Overplacement Paradigm Validity

The connections between many of the phenomena discussed in Section 5.1 and overconfidence have largely been established through argumentation and axiomatic reasoning rather than empirical evidence, as note by Merkle and Weber [2011]. Consequently, the apparent robustness of overconfident relative self-assessments as a stable psychological or behavioral pattern - whether in the early assumption that individuals are generally overconfident [DeBondt and Thaler, 1995; Alicke et al., 1995] or in the more nuanced perspective that individuals tend to overplace themselves on easy tasks while underplacing on difficult ones [Moore and Healy, 2008; Kruger et al., 2008; Grieco and Hogarth, 2009] - may be questioned [Benoît and Dubra, 2011; Benoît et al., 2014; Owens et al., 2014]. More specifically, it has been argued that "better-than-average" data may not be able to demonstrate overconfidence [Benoît and Dubra, 2011], and that inferring overconfident beliefs from choice behavior likely leads to overestimation of the actual extent of overconfidence [Owens et al., 2014].

In particular, Benoît and Dubra [2011] challenge the axiomatically established presumption in psychology and economics that a rational population should exhibit neither overplacement nor underplacement. They argue that the statement *"most people cannot be better than the median"* does not logically imply that *"most people cannot rationally rate themselves above the median"* (p. 1592). Individuals may have valid reasons to believe they will perform in the upper half of a population. Therefore, misplacement in self-assessments is *"unproblematic if [it]] can arise from a population whose beliefs are generated within a rationalizing model"* (p. 1594). According to their framework, this holds for almost any distribution of placement relative to the median. Thus, what appears to be overplacement may actually be consistent with rational Bayesian updating, given that individuals lack perfect knowledge of their own abilities and the overall distribution of skills in a population [Benoît and Dubra, 2011; Benoît et al., 2014; Moore and Healy, 2008]. Under conditions of perfect information, only 50% of individuals could, in fact, rationally place themselves in the top half of a given population.

With imperfect information, however, for a population to be truly considered overconfident, substantially more than 50% would need to rate themselves above the median. Mathematically, Benoît and Dubra [2011] derive that up to twice the proportion of any given percentile may rationally rate themselves within the respective top percentage of a population (e.g., 60% believing they are in the top 30%) and propose this as a threshold for inferring overconfidence from overplacement data. Consequently, practically an entire population could rationally expect to be in its top half.

Accordingly, ranking experiments that rely on a single reference point - typically the median - and

employ a dichotomous categorization of success to infer miscalibration in relative abilities generally reveal only 'apparent', rather than 'true', overconfidence or underconfidence. TMany experimental findings of overplacement may therefore be median-rationalizable within Benoît and Dubra [2011]'s theoretical framework, as demonstrated by Benoît et al. [2014] in their reassessment of results from Hoelzl and Rustichini [2005] and Moore and Cain [2007]. This distinction is crucial because true overconfidence can lead to the negative real-world consequences outlined in Section 5.1, whereas apparent overconfidence is unlikely to have such effects [Benoît et al., 2014].

Methodologically, Benoît and Dubra [2011] criticize commonly used incentive-compatible probability elicitation measures, such as quadratic scoring rules, for often being unintuitive and failing to penalize incorrect estimations. Additionally, key aspects of individuals' subjective understanding of the skill or ability in questions - such as their interpretation of "average" (whether as mean, median, or mode), response scales, and the signaling structure shaping their beliefs - are typically unobservable, compromising both internal validity within studies and comparability across studies[50] [Benoît and Dubra, 2011; Benoît et al., 2014]. Therefore, Benoît and Dubra [2011] call for experimental designs that collect more detailed information on the strength of subjects' beliefs beyond rankings relative to the median - for instance, by asking participants to indicate their perceived likelihood of placing within specific deciles - to be able to actually study observe overconfidence. Expanding on this, Benoît et al. [2014] advocate for a more nuanced interpretation of individuals' self-assessments relative to the median. Subjects favoring placement-based rewards should not necessarily be interpreted as these individuals believing they are top-half performers with certainty - which would indicate true overconfidence - but rather as them assuming to have at least a 50% chance of placing in the top half. This belief, they contend, is consistent with rational expectations that could be held by nearly every individual in a population. In two experiments involving easy quizzes with complete information on prior performances, Benoît et al. [2014] find overplacement data that are not rationalizable for a population of expected payoff maximizers under Benoît and Dubra [2011]'s framework, and, therefore, appear to display true overconfidence. Their explanation of individuals extrapolating good absolute performance to above average relative performance while neglecting competitors' performance goes along with findings from prior research [e.g., Kruger, 1999; Alicke and Govorun, 2005; Kruger et al., 2008; Moore and Small, 2007], and matches correlational patterns observed in this paper (see Section 5.5).

Several studies argue that overplacement should be considered an information or statistical bias

[50]For example, in Svenson [1981]'s prominent experiment on driving skill ranking - besides it being unincentivized - participants lacked objective criteria for accurately assessing their driving ability due to limited available information. Applying Benoît and Dubra [2011]'s criteria, the reported proportion of better-than-median assessments suggests that the Swedish population's responses can be median-rationalized, whereas the American population cannot. Consequently, Svenson identifies 'true' overconfidence in the American population but only 'apparent' overconfidence in the Swedish.

rather than a confidence or behavioral bias, stemming from incomplete information on the respective reference group [Healy and Moore, 2007; Moore and Small, 2007; Blavatskyy, 2009], or imperfect Bayesian updating to information about others [Moore and Healy, 2008; Grieco and Hogarth, 2009; Grossman and Owens, 2012]. Symptomatically, the tendency to rate oneself as above average on easy tasks and below average on difficult tasks becomes stronger if an individual receives information about their performance and is attenuated if they receive information about a reference group's performance [Moore and Small, 2007]. Moreover, incomplete information on a task's difficulty or the distribution of relevant skills within the reference population may lead to either overconfident or underconfident beliefs [Healy and Moore, 2007; Blavatskyy, 2009]. Especially, simultaneous inference about task difficulty and relative performance may lead to mixed and strongly context-dependent evidence on misplacement in either direction [Benoît et al., 2014; Grieco and Hogarth, 2009; Clark and Friesen, 2009; Moore and Healy, 2008]. Therefore, the statistical impossibility of more than half a population assessing themselves in its top half does not necessarily indicate an underlying judgment bias, i.e., overconfident beliefs [Healy and Moore, 2007].

Beyond the discussion on thresholds for justifiably inferring overconfidence from overplacement data and the role of information availability in shaping miscalibrated self-assessments, there may simply be additional non-monetary motives influencing individuals' placement behavior, which are difficult to disentangle from truely overconfident beliefs in discrete choice experiments. If overplacement in choice behavior is rooted in (true) overconfidence, i.e., a psychological misperceptions or errors in beliefs, better information could lead to adjusted behavior. Conversely, if overplacement is driven by alternative motives - such as control preferences, self-enhancement, image concerns, or attitudes toward risk and ambiguity - behavior is unlikely to change [Benoît et al., 2014; Owens et al., 2014; Benoît et al., 2022].

For instance, Heath and Tversky [1991] argue that humans exhibit a distinct preference for betting on themselves over a random device - even when the latter offers an equal or higher success probability - provided they feel competent in the given context. Beyond that, Goodie and Young [2007] observe individuals to display a general preference for control across various domains. Owens et al. [2014] show that individuals sacrifice up to 15% of expected earnings to bet on themselves — a "control premium"[51] that cannot be fully explained by beliefs. Instead of offering participants the choice between betting on their own performance and a random device - as in H&R - they let subjects choose between betting on themselves or on someone else. Their results provide an estimate of more than half (57%) of the observed self-reliance beyond 50% stemming from a preference for control (64.9%

[51] "Control premium" is commonly used as an umbrella term that encompasses multiple reasons why individuals prefer to bet on themselves. These include a general preference for autonomy in decision-making, greater enjoyment of betting on oneself, the desire to signal competence, or a broader inclination toward self-reliance in wagering [Owens et al., 2014; Benoît et al., 2022].

chose to bet on themselves overall), leaving the remainder attributable to overconfident beliefs [Owens et al., 2014]. Using these reference values to reinterpret H&R's results, out of the 63 percent self-reliance in the *Easy x Money* treatment, only 5.5 percentage-points of the deviation from 50 percent could be attributed to overconfidence, while the remaining 7.5 percentage-points would be explained by a preference for control. However, it must be noted that a desire for control likely would not explain underplacement data, apart from individuals exhibiting a (strong) negative control premium. Owens et al. [2014] do neither explicitly comment on the *Difficult x Money* condition from H&R nor underplacement in general. Benoît et al. [2022] replicate Owens et al. [2014]'s experimental design with an adapted treatment condition in which participants choose between betting on their performance in one task versus another, ensuring individuals bet on themselves either way to mitigate potential control distortions related to ambiguity aversion towards another person's abilities. They also compare betting on one's own performance to betting on a random device with an identical probability of success. Their results suggest that at least 68% of the apparent overconfidence observed in participants' choices between betting on their own performance and a random device (14.2 percentage-points in total) stemming from control-related preferences. When applying their adapted mechanism to mitigate control preferences, Benoît et al. [2022] still observe a small but statistically significant degree of overplacement (54% of participants expect to rank in the top 50%), indicating the true overconfidence.

Merkle and Weber [2011] directly address Benoît & Dubra's critique through two experiments that meet their criteria, by eliciting participants' complete belief distributions regarding their relative placement within a population across multiple domains and tasks - capturing probability estimates for each decile or quartile rather than relying on point estimates - to distinguish between rational information processing and (true) overconfidence as competing explanations for overplacement. Findings reveal considerable overplacement across several domains, while displaying underplacement for high specialization abilities (e.g., programming skills) and low-probability events (e.g., likelihood of suffering a heart attack), with probability estimates aligning with prior research based on point estimates. Rational information processing is rejected for four out of six domains, while results for the remaining two are mixed. Consistent with Benoît et al. [2014], Merkle & Weber find that overplacement in aggregated belief distributions is especially driven by low-skilled individuals strongly underperforming relative to already poor self-assessments, despite receiving negative signals.

Ultimately, Merkle and Weber [2011, p.262] reaffirm that *"overconfidence is not just an artifact of psychological experiments, but seems present in many real-life situations where considerable stakes are involved"*. While acknowledging the theoretical validity of Benoît and Dubra's critique, they argue that its practical consequences may be limited, as overplacement persists even after accounting for rational adjustments. Accordingly, they conclude that *"there is no need to discard previous literature*

on this premise" (p. 271), though they support incorporating the discussed methodological refinements into experimental designs [Merkle and Weber, 2011]. Benoît et al. [2014] oppose this by stating that, despite some evidence withstanding the criticism by Benoît and Dubra [2011], the actual scope and significance of overplacement in human judgment and decision-making remains unclear, limited number of studies employing rigorous experimental designs. Furthermore, they reemphasize that *"in any case, it is important to realize that the degree of (true) overconfidence may not be well measured by the fraction of people who rank themselves as above average"* (p. 321).

D.2 Additional Tables and Figures

Table 45: Demographic Statistics

Variable	Total	Replication (unbalanced)	Adaptation	Replication (balanced)
Number of observations	272	90	92	90
Age	34.3	31.2	34.1	37.7
(SD)	(11.2)	(9.4)	(11.6)	(11.9)
Female	44.1	30.0	50.0	52.2
Student	32.4	38.9	37.0	21.1
Study major or professional field				
Business, Economics, Marketing, Personnel, Sales, or Insurance	19.5	18.9	18.5	21.1
Design, Communications, or Media	4.0	2.2	6.5	3.3
Education, Cultural Studies or Public Sector	15.44	7.8	17.4	21.1
Engineering, Manufacturing, Construction, or Agriculture	10.3	8.9	14.1	7.8
Information Technology and Natural Sciences	18.4	24.4	15.2	15.6
Medicine, Psychology, Health, or Social Care	8.5	10.0	7.6	7.8
Social Studies, Journalism, or Politics	4.0	5.6	5.4	1.11
Transport, Logistics, Retail, or Wholesale	5.5	6.7	6.5	3.33
Homemaker, Unemployed, Retired, Pupil, or not specified	14.0	14.4	8.7	18.9

Note: Averages reported for Age. Relative frequencies reported for all other variables.

Table 46: Subjects' Self- and Group Assessments Before and After the Test & Questionnaire Control Summary Statistics, by Treatment

	Replication		Adaptation		Mann-Whitney U	
Variable	Mean	Std. Dev.	Mean	Std. Dev.	z	p-value
Before						
Predicted own performance	13.54	3.20	14.28	3.60	1.77	0.0766
Predicted group performance	12.27	2.40	13.47	2.27	3.18	0.0014
Better	1.118	0.25	1.079	0.28	-0.32	0.7530
Before						
Sample performance	2.36	0.85	2.41	0.74	0.18	0.8593
Test performance	13.6	3.1	13.1	3.3	-0.83	0.4074
After						
Estimated group performance	12.6	2.6	12.6	2.7	-0.13	0.8988
Estimation accuracy	-0.1	3.5	1.2	4.3	1.93	0.0530
Questionnaire Controls						
INCOM	3.56	0.66	3.54	0.69	0.22	0.8297
GSE	3.84	0.77	3.91	0.57	0.03	0.9810
Altruism	3.77	0.72	3.84	0.70	-0.79	0.4308
Risk (MPL)	15.35	6.43	15.39	5.63	-0.22	0.8269
Risk (11-point)	6.13	2.47	5.58	2.33	1.489	0.1368
Ambiguity	12.58	5.86	12.16	5.43	0.416	0.6788

Note: INCOM: Short scale of Iowa–Netherlands Comparison Orientation Measure (5-point scale); *GSE*: General self-efficacy (5-point scale); *Altruism*: Altruism facet from HEXACO-100 (5-point scale); *Risk (MPL)*: Multiple price list format for measuring risk preferences; *Risk (11-point)*: General willingness to take risk (11-point scale); *Ambiguity*: Multiple price list format for measuring ambiguity aversion. "Mann-Whitney U" reports results for Two-sample Mann–Whitney U-tests between treatments. Replication n = 90; Adaptation n = 92.

Table 47: Regression of Actual Performance over Predicted Own Performance

	R^2	F	Prob $> F$
	0.0501	13.66	0.0003
Actual own performance	Coefficient	t	$P > t$
Predicted own performance	0.2201 (0.0595)	3.70	0.000
Constant	9.7702 (0.8420)	11.60	0.000

Note: Refers to Table 16 in H&R (Technical Appendix).

Table 48: Summary Statistics, Before the Test

	Full sample		H&R		Two-sample t-test	
Variable	Mean	Std. Error	Mean	Std. Error	t	p-value
Predicted own performance	13.94	0.22	12.51	0.30	3.85	0.0001
Predicted group performance	12.99	0.16	13.37	0.24	-1.36	0.1751
Sure	5.29	0.08	5.16	0.14	0.85	0.3964
Difficult to change	4.23	0.09	3.66	0.15	3.33	0.0009
Important	5.96	1.32	3.24	0.16	17.54	0.0000
Difficulty test	5.46	0.06	4.58	0.10	7.62	0.0000
Good in test	4.91	0.07	4.29	0.10	4.92	0.0000

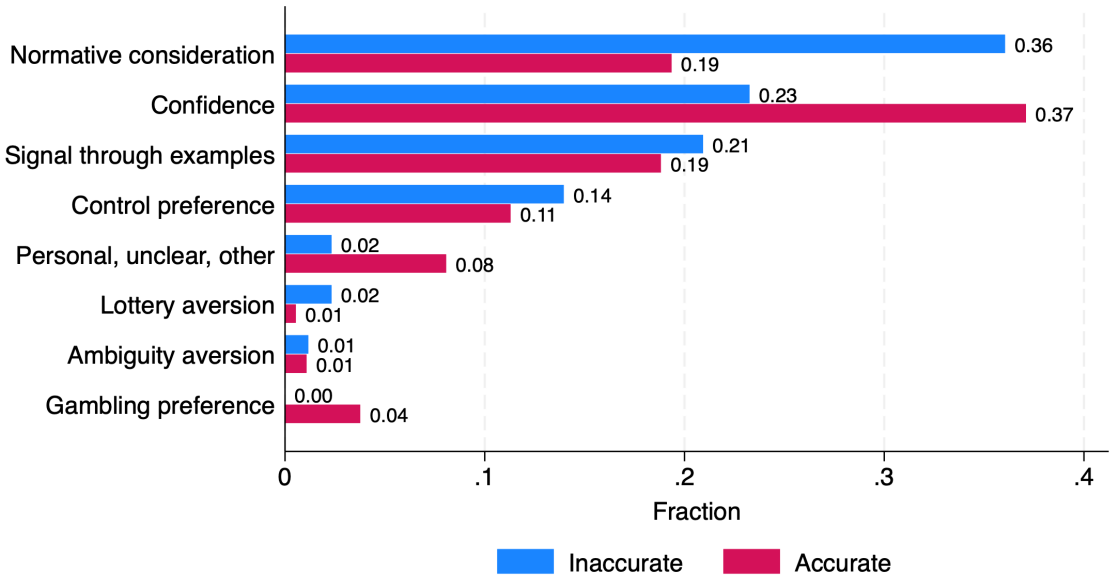
Note: sure: "How sure are you to have made the right decision in the vote?" (7 = very sure); *difficult to change*: "How difficult would you find it to change your decision in the vote?" (7 = very difficult); *important*: "How important is doing well in the test to you?" (7 = very important); *difficulty test*: "How difficult do you think the test will be?" (7 = very difficult); *good in test*: "How good do you think you will be in the test?" (7 = very difficult); N = 272; H&R n = 134. Since H&R report standard errors of the mean instead of standard deviations, this is adopted for comparability. Right column gives results from two-sided Two-sample t-tests with equal variances.

Table 49: Summary Statistics, After the Test

	Full sample		H&R		Two-sample t-test	
Variable	Mean	Std. Error	Mean	Std. Error	t	p-value
Actual own performance	12.84	0.22	11.83	0.45	2.29	0.0223
Estimated group performance	12.52	0.17	13.43	0.30	-2.86	0.0045
Satisfied	4.11	0.11	4.34	0.18	-1.12	0.2624
Sure after	4.88	0.10	5.33	0.15	-2.47	0.0139
Difficult to change after	4.42	0.11	4.15	0.18	1.31	0.1899
Difficulty test after	5.66	0.06	4.48	0.14	8.82	0.0000

Note: satisfied: "How satisfied are you with your result in the test?" (7 = very sure); *sure after*: "How sure are you now to have made the right decision in the vote?" (7 = very difficult); *difficult to change after*: "How difficult would you find it now to change your decision in the vote?" (7 = very important); *difficulty test after*: "How difficult do you think the test was?" (7 = very difficult); N = 272; H&R n = 134. Since H&R report standard errors of the mean instead of standard deviations, this is adopted for comparability. Right column gives results from two-sided Two-sample t-tests with equal variances.

Figure 27: Voting Rationales by Accuracy in Calibration, Self-reported by Subjects



Note: Accurate: n = 186, Inaccurate: n = 86.

Table 50: Logit Regression for Voting Behavior – Coefficients

	Dependent variable: Likelihood of Vote for 'Test'				
	(1)	(2)	(3)	(4)	(5)
Intercept	-1.259	-1.829	-4.004	2.655	-1.613
	(0.905)	(0.925)	(1.796)	(1.640)	(2.296)
Own Performance Prediction	0.400	0.406	0.411	0.471	0.473
	(0.070)	(0.072)	(0.090)	(0.077)	(0.103)
Group Performance Prediction	-0.382	-0.384	-0.385	-0.408	-0.431
	(0.084)	(0.084)	(.093)	(.082)	(0.101)
Sample Performance	0.775	1.041	0.858	0.685	1.148
	(0.231)	(0.227)	(0.246)	(0.241)	(0.275)
Age		0.016			0.014
		(0.024)			(0.026)
Age^2		0.002			0.002
		(0.001)			(0.001)
Female		-1.372			-1.644
		(0.376)			(0.456)
Student		1.038			1.418
		(0.444)			(0.490)
Degree in Higher Education		0.181			0.604
		(0.366)			(0.422)
Sure to have made the right Vote			-0.128		-0.067
			(0.132)		(0.159)
Difficulty to change Vote			0.044		0.021
			(0.115)		(0.127)
Importance of doing well in Test			0.115		0.257
			(0.125)		(0.166)
Expected Test Difficulty			0.303		0.354
			(0.179)		(0.227)
Expectation of doing well in Test			0.124		0.274
			(0.186)		(0.203)
Social Comparison Orientation (INCOM)				-0.534	-0.583
				(0.242)	(0.273)
General Self-Efficacy				-0.407	-0.622
				(0.261)	(0.353)
Altruism				-0.168	-0.168
				(0.280)	(0.371)
Risk Attitude				-0.067	-0.121
				(0.075)	(0.092)
Ambiguity Attitude				0.009	0.021
				(0.030)	(0.034)
Wald $\chi^2(3)$	47.72	76.31	49.50	53.71	62.29
Pseudo R^2	0.291	0.361	0.309	0.321	0.417
N	272	272	272	272	272

Note: Coefficients estimated using robust standard errors, standard errors in parentheses.
Model specifications: (1) Base model (see Table 25), (2) including demographics, (3) including pre-task expectations, (4) including questionnaire controls, (5) full model. Only the single-item 11-point risk measure by Dohmen et al. [2011] is included for simplicity. Study major and professional field not included due to arbitrary categorization.

Table 51: Logit Regression for Voting Behavior – Marginal effects

	Dependent variable: Likelihood of Vote for 'Test'				
	(1)	(2)	(3)	(4)	(5)
Own Performance Prediction	0.054***	0.049***	0.054***	0.061***	0.051***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Group Performance Prediction	-0.052***	-0.047***	-0.051***	-0.053***	-0.046***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Sample Performance	0.105***	0.127***	0.113***	0.089**	0.126***
	(0.000)	(0.000)	(0.000)	(0.003)	(0.000)
Age		0.002			0.002
		(0.460)			(0.490)
Female		-0.167***			-0.180***
		(0.000)			(0.000)
Student		0.126*			0.155**
		(0.020)			(0.003)
Degree in Higher Education		0.022			0.066
		(0.620)			(0.146)
Sure to have made the right Vote			-0.017		-0.007
			(0.323)		(0.671)
Difficulty to change Vote			0.006		0.002
			(0.698)		(0.866)
Importance of doing well in Test			0.015		0.028
			(0.358)		(0.122)
Expected Test Difficulty			0.040		0.038
			(0.090)		(0.112)
Expectation of doing well in Test			0.016		0.030
			(0.507)		(0.169)
Social Comparison Orientation (INCOM)				-0.069*	-0.064*
				(0.021)	(0.029)
General Self-Efficacy				-0.053	-0.068
				(0.110)	(0.059)
Altruism				-0.022	-0.018
				(0.548)	(0.652)
Risk Attitude				-0.009	-0.013
				(0.364)	(0.183)
Ambiguity Attitude				0.001	0.002
				(0.757)	(0.528)

Note: p-values in parentheses; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Table 52: Aspect by Aspect Comparison of Experimental Designs

Hoelzl & Rustichini (2005)	Protte (2025)
Environment	
• Laboratory	• Remote
• Paper & pen	• Computerized
Recruitment	
• University campus	• Prolific.com
Sample	
• N = 134	• N = 200 (182 in analysis)
• Mostly university students (plus 6 professionals, 4 pupils)	• Heterogeneous sample (32% students)
• ø age: 23	• ø age: 34.3
Treatments	
• Incentive (money vs. no money)	• Lottery mechanism (probabilistic outcome distribution vs. fixed outcome distribution)
• Task difficulty (hard vs. easy)	
Session sizes	
• Multiple groups of "at least 9" subjects, one group of 7	• Two groups of 100 subjects
Indicator variable	
• Vote on payoff mechanism	• Vote on payoff mechanism
Task	
• LEWITE (20 items)	• SAT Analogies (20 items)
• Choose 2 from 7-9 alternatives	• Choose 1 from 5 alternatives
Payment	
• Fixed payment: None	Fixed payment: 3.00 EUR ($\sim$ 3.30 USD)
• Bonus: 150 ATS ($\tilde{1}$0 USD at the time)	Bonus: 5.00 EUR ($\sim$ 5.50 USD)
Procedure	
• Instructions distributed and read out aloud	• Instructions presented on screen
	• Comprehension questions
• Practice examples and solutions	• Practice examples and solutions
• Vote	• Vote & Voting rationale
• Pre-task questions and predictions	• Pre-task questions and predictions
• Test (results revealed to subjects but otherwise kept confidential)	• Test (results revealed to subjects but otherwise kept confidential)
• Individual die roll	• Post-task questions and estimations
• Post-task questions and estimations	• Choice of lucky number (die roll equivalent)
	• Additional questionnaires (INCOM, GSE, HEXACO altruism, MPL risk, MPL ambiguity, demographics)
• Voting result & median test result announced verbally	• Voting result & median test result announced via bulk mail

D.3 Experiment Instructions

[Translated from German]

Welcome!

Thank you for participating in today's study!

This is a non-commercial study conducted by researchers at Paderborn University and funded by Paderborn University.

Please note the following information:

- The data collected through this study is completely anonymous and is used exclusively for scientific purposes. It is not possible to draw conclusions about you or other persons.
- Your data will be conscientiously protected in accordance with the provisions of the European General Data Protection Regulation (GDPR), even in the case of a scientific publication.
- Your participation in the study is voluntary and can be terminated at any time without giving reasons.

Please answer the questions as carefully and conscientiously as possible. Please only start working on the study when you have enough time to answer them in one sitting. Communication programs (e.g. chat or e-mail) should be closed during processing in order to avoid distractions.

Please press "Start experiment" to acknowledge that you have read and agreed with the conditions as stated above. If you do not agree with the conditions, please close your browser window.

***

Instructions

Please read the following instructions carefully, as you will be asked comprehension questions on the next page!

You will receive a fixed payment of £2.00 for your participation. In addition, you can receive a bonus payment of £5.00. The conditions for this bonus payment are explained below.

Not all participants will receive a bonus payment. Whether you receive a bonus payment is decided either by (a) a performance test or (b) a lottery.

In order to decide which payoff mechanism (performance test or lottery) determines whether you receive a bonus payment, all participants in this experiment will vote on the payoff mechanism at the beginning:

- Under the "performance test" condition, you will receive a bonus payment if your score is in the upper half of all 100 participants, i.e. if your performance is among the top 50 participants.
- *[Replication group]* Under the "Lottery" condition, you choose a lucky number between 1 and 6. After all 100 people have taken part, a random generator throws a conventional 6-sided dice. If the dice shows an odd number (i.e. 1, 3, 5), you will receive a bonus payment if you have entered an odd number as your lucky number. If the dice shows an even number (i.e. 2, 4, 6), you will receive a bonus payment if you have entered an even number as your lucky number.
- *[Adaptation group]* Under the "Lottery" condition, you choose a lucky number between 1 and 6, which, together with your Prolific ID, forms your winning code. After all 100 people have taken part, a random generator draws 50 winning codes. If your winning code is drawn, you will receive a bonus payment.

In total, 100 people will participated in this study. A simple majority will decide by secret vote which payoff mechanism will be applied.

If the majority vote in favor of the "performance test" payoff mechanism, your test result will determine whether you receive a bonus payment. If the majority of participants vote in favor of the "lottery" payoff mechanism, the lottery result will decide whether you receive a bonus payment. (If there is an exact tie, a coin toss by the experimenters will decide).

As you do not yet know at the time of your participation which payout mechanism has ultimately been voted, you will play both the test and the lottery. Only you will be informed of your individual test result.

As soon as all 100 people have taken part in this study, you will be informed via an email to your Prolific ID about which payment mechanism has been voted and whether you will receive a bonus payment. You will also be informed about the average test performance of all participants.

Please make sure that you have read and understood the instructions, as you will be asked some comprehension questions on the following page. Following the comprehension questions, you will receive three examples of the test tasks.

After the experiment section, we will ask you to complete questionnaire in which there are no "right" or "wrong" answers. Please simply answer the questions according to your personal judgments and opinions.

(The "Next" button appears after 90 seconds.)

***

Comprehension Questions

[Question 5 differed between treatments, 'X' indicates correct answer]

1) Which of the following statements about the procedure of the experiment is correct?

- ○ You will play both the test and the lottery, regardless of which you choose in the vote. X
- ○ You will either only play the test or only the lottery and decide in the vote.

2) Which of the following statements about the payoff mechanism is correct?

- ○ You alone decide which payout mechanism is used for you.
- ○ A majority vote among all participants determines which payout mechanism will be used for all participants. X

3) Which of the following statements about the bonus payment is correct?

- ○ All participants receive a bonus payment.
- ○ The test result or the lottery result (depending on the voting result) decides who receives a bonus payment. X

4) Which of the following statements about the performance test is correct?

- ○ If the performance test is voted as the payoff mechanism, 50 randomly selected participants will receive a bonus payment.
- ○ If the performance test is voted as the payoff mechanism, half of the participants (the 50 with the best test results) will receive a bonus payment. X

5 - Replication group) Which of the following statements about the lottery is correct?

- ○ If the lottery is voted as the payoff mechanism, you have a 50% probability of receiving a bonus payment (namely if you have selected an odd lucky number and an odd number is rolled or if you have selected an even lucky number and an even number is rolled). X
- ○ If the lottery is voted as the payout mechanism, the probability of not receiving a bonus payment is higher than the probability of receiving a bonus payment.

5 - Adaptation group) Which of the following statements about the lottery is correct?

- ○ If the lottery is voted as the payoff mechanism, exactly half of the participants (50 out of 100) will receive a bonus payment (i.e. if their winning code is drawn). X
- ○ If the lottery is voted as the payoff mechanism, more participants will receive a bonus payment than in the performance test.

6) When will you be informed about your test result?

○ Immediately after completing the test. X
○ After all 100 people participated in the study.

7) When will you be informed whether you will receive a bonus payment?

○ Immediately after completing the test.
○ After all 100 people participated in the study. X

***

Vote

Your task in the test is to identify the most appropriate analogy between pairs of words. Below you can see three examples of test tasks.

In total, the test consists of **20 tasks**.

Please select one of the five possible answers each. The correct solution will be displayed on the next page.

Which pair of words relates to each other like:

1) "seed" to "plant"

○ pouch - kangaroo ○ root - soil ○ drop - water ○ bark - tree ○ egg - bird

2) "unfetter" to "pinioned"

○ recite - practiced ○ sully - impure ○ enlighten - ignorant ○ revere - unrecognized ○ adore - cordial

3) "vacillate" to "indecision"

○ lament - woe ○ hibernate - winter ○ extricate - entanglements ○ digress - angst ○ emulate - egotism

[Correct solutions underlined here for conciseness.]

***

Now, please cast your vote for the choice of payoff mechanism and then click "Confirm".

Your vote:

○ Test ○ Lottery

(The "Next" button appears after 30 seconds.)

***

Please explain the rationale for your choice in 1-2 sentences: _____

***

Performance test

Please indicate your expectations for the following aspects:

How sure are you that you have made the right decision in the vote?

Very unsure ○ ○ ○ ○ ○ ○ ○ Very sure

How difficult would you find it to change your decision?

Very easy ○ ○ ○ ○ ○ ○ ○ Very difficult

How important is doing well in the test for you?

Not important at all ○ ○ ○ ○ ○ ○ ○ Very important

How difficult do you expect the test to be?

Very easy ○ ○ ○ ○ ○ ○ ○ Very difficult

How do you think you will do in the test?

Very bad ○ ○ ○ ○ ○ ○ ○ Very good

Please specify:

How many of the 20 questions will **you** answer correctly?

How many of the 20 questions will **the other participants** answer correctly **on average**?

***

Thank you for your participation on this experiment so far!

The announced test will begin on the following page.

Please only advance when you have enough time to complete the tasks.

The test consists of 20 tasks in total. You have 30 seconds per task.

At the top of the screen you will see a timer that starts automatically as soon as you begin the test.

***

Please select one of the five possible answers and then click "Next".

Which pair of words relates to each other like:

1) "emulate" - "person"

○ admire - reputation ○ obey - leader ○ cooperate - partner ○ mimic - gesture ○ mock - sarcasm

2) "irrational" - "logic"

○ unrealistic - understanding ○ unethical - morality ○ illegible - erasure ○ infinite - expansion ○ factual - verification

3) "paint" - "canvas"

○ brush - bucket ○ peanut - butter ○ clock - time ○ art - life ○ no - choice

4) "mansion" - "residence"

○ limousine - automobile ○ chandelier - candle ○ tuxedo - wardrobe ○ diamond - rhinestone ○ yacht - harbor

5) "resolute" - "determination"

○ pristine - grace ○ skeptical - doubt ○ tainted - honor ○ stringent - suggestions ○ wary - risks

6) "actor" - "cast"

○ musician - orchestra ○ singer - song ○ lecturer - class ○ congregation - church ○ proofreader - text

7) "library" - "book"

○ scholar - knowledge ○ stable - horse ○ factory - outlet ○ laboratory - radiation ○ laser - energy

8) "emissary" - "represent"

○ draftee - enroll ○ novice - train ○ president - elect ○ guard - protect ○ comedian - laugh

9) "format" - "newspaper"

○ binding - book ○ design - building ○ direction - sign ○ market - commodity ○ catalog - library

10) "liter" - "volume"

○ day - night ○ mile - distance ○ decade - century ○ friction - heat ○ part - whole

11) "song" - "repertoire"

○ score - melody ○ instrument - artist ○ solo - chorus ○ benediction - church ○ suit - wardrobe

12) "magnify" - "appearance"

○ rectify - error ○ augment - reduction ○ probe - feeling ○ amplify - volume ○ distort - image

13) "canal" - "waterway"

○ skyline - city ○ bank - stream ○ hub - wheel ○ dam - river ○ reservoir - lake

14) "incorrigible" - "reformed"

○ unnerving - irritated ○ innocuous - harmed ○ irrelevant - verified ○ insolvent - dissolved ○ indelible - erased

15) "hostile" - "bellicose"

○ indifferent - averse ○ stubborn - obdurate ○ morose - slothful ○ unequivocal - skeptical ○ angry - passive

16) "border" - "country"

○ current - river ○ water - lake ○ waves - sea ○ horizon - sunset ○ shore - ocean

17) "jubilation" - "joy"

○ exaggeration - truth ○ compassion - sympathy ○ security - instability ○ fortitude - danger ○ emotion - anger

18) "diversion" - "boredom"

○ assurance - uncertainty ○ enmity - hatred ○ secrecy - curiosity ○ reward - deed ○ sluggishness - fatigue

19) "sprout" - "seed"

○ pollinate - bee ○ cure - disease ○ stimulate - growth ○ hatch - egg ○ filter - impurity

20) "prudent" - "indiscretion"

○ frugal - wastefulness ○ proud - accomplishment ○ generous - wealth ○ disqualified - competition ○ disgruntled - cynicism

***

Please specify:

How many of the 20 questions did **the other participants** answer correctly **on average**?

In retrospect, please assess the following aspects:

How satisfied are you with your test result?

Not satisfied at all ○ ○ ○ ○ ○ ○ ○ Very satisfied

How sure are you now that you have made the right decision in the vote?

Very unsure ○ ○ ○ ○ ○ ○ ○ Very sure

How difficult would you find it now to change your decision?

Very easy ○ ○ ○ ○ ○ ○ ○ Very difficult

How difficult did you find the test?

Very easy ○ ○ ○ ○ ○ ○ ○ Very difficult

***

Lottery

You have now completed the test. Thank you for participating in the experiment so far!

On the following page we will continue with the lottery.

***

[Lottery Replication]

Reminder:

In this lottery, you choose a lucky number between 1 and 6. After all 100 people have taken part, a random number generator throws a conventional **6-sided dice**.

If the dice shows an odd number (i.e. 1, 3, 5), you will receive a bonus payment if you have entered an odd number as your lucky number.
If the dice shows an even number (i.e. 2, 4, 6), you will receive a bonus payment if you have entered an even number as your lucky number.

Please enter your lucky number here: ___

[Lottery Adaption]

Reminder:

In this lottery, you choose a lucky number between 1 and 6. This, together with your Prolific-ID, forms your **winning code**.

After all 100 people have taken part, a random generator draws 50 winning codes. If your winning code is drawn, you will receive a bonus payment.

Please enter your lucky number here: ___

***

Questionnaire

Thank you for completing the experiment so far!

In the following, we ask you to complete a multi-part questionnaire. There are no "right" or "wrong" answers. Simply indicate to what extent the following statements apply to you personally.

***

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
I always pay a lot of attention to how I do things compared with how others do things.	○	○	○	○	○
I often compare how I am doing socially (e.g., social skills, popularity) with other people.	○	○	○	○	○
I am not the type of person who compares often with others.	○	○	○	○	○
Please choose "Strongly disagree".	○	○	○	○	○
I often try to find out what others think who face similar problems as I face.	○	○	○	○	○
I always like to know what others in a similar situation would do.	○	○	○	○	○
If I want to learn more about something, I try to find out what others think about it.	○	○	○	○	○

***

	Does not apply at all	Applies little	Somewhat applies	Pretty much applies	Fully applies
In difficult situations I can rely on my abilities.	○	○	○	○	○
I can cope well with most of the problems on my own power.	○	○	○	○	○
Even strenuous and complicated tasks I can usually solve well.	○	○	○	○	○

***

	Strongly agree	Agree	Neither degree nor disagree	Disagree	Strongly disagree
I have sympathy for people who are less fortunate than I am.	○	○	○	○	○
I try to give generously to those in need.	○	○	○	○	○
It wouldn't bother me to harm someone I didn't like.	○	○	○	○	○
People see me as a hard-hearted person.	○	○	○	○	○

***

Please see the following instructions.

The following table contains 29 separate decisions.

Please decide for each row which lottery you prefer: "Option A" or "Option B".

To specify your preferences between Option A and Option B, you only need to make one choice, namely in which row you would like to switch **from Option A to Option B**.

For each row **before** the one you have selected, you prefer Option A. For each row **after** the one you have selected, including the one you have selected, you prefer Option B.

	Option A	**Option B**
1)	€30 with a probability of 50%, €0 with a probability of 50%	€1 safe
2)	€30 with a probability of 50%, €0 with a probability of 50%	€2 safe
3)	€30 with a probability of 50%, €0 with a probability of 50%	€3 safe
4)	€30 with a probability of 50%, €0 with a probability of 50%	€4 safe
5)	€30 with a probability of 50%, €0 with a probability of 50%	€5 safe
6)	€30 with a probability of 50%, €0 with a probability of 50%	€6 safe
7)	€30 with a probability of 50%, €0 with a probability of 50%	€7 safe
8)	€30 with a probability of 50%, €0 with a probability of 50%	€8 safe
9)	€30 with a probability of 50%, €0 with a probability of 50%	€9 safe
10)	€30 with a probability of 50%, €0 with a probability of 50%	€10 safe
11)	€30 with a probability of 50%, €0 with a probability of 50%	€11 safe
12)	€30 with a probability of 50%, €0 with a probability of 50%	€12 safe
13)	€30 with a probability of 50%, €0 with a probability of 50%	€13 safe
14)	€30 with a probability of 50%, €0 with a probability of 50%	€14 safe
15)	€30 with a probability of 50%, €0 with a probability of 50%	€15 safe
16)	€30 with a probability of 50%, €0 with a probability of 50%	€16 safe
17)	€30 with a probability of 50%, €0 with a probability of 50%	€17 safe
18)	€30 with a probability of 50%, €0 with a probability of 50%	€18 safe
19)	€30 with a probability of 50%, €0 with a probability of 50%	€19 safe
20)	€30 with a probability of 50%, €0 with a probability of 50%	€20 safe
21)	€30 with a probability of 50%, €0 with a probability of 50%	€21 safe
22)	€30 with a probability of 50%, €0 with a probability of 50%	€22 safe
23)	€30 with a probability of 50%, €0 with a probability of 50%	€23 safe
24)	€30 with a probability of 50%, €0 with a probability of 50%	€24 safe
25)	€30 with a probability of 50%, €0 with a probability of 50%	€25 safe
26)	€30 with a probability of 50%, €0 with a probability of 50%	€26 safe
27)	€30 with a probability of 50%, €0 with a probability of 50%	€27 safe
28)	€30 with a probability of 50%, €0 with a probability of 50%	€28 safe
29)	€30 with a probability of 50%, €0 with a probability of 50%	€29 safe

Please indicate the row for which you want to switch from Option A to Option B: ___

(The "Next" button appears after 60 seconds.)

***

Please see the following instructions.

The following table contains 20 separate decisions.

First, please choose your success color: ○ **Red** ○ **Black**

Afterwards, please decide for each row which lottery you prefer: "Urn A" or "Urn B". Both urns contain exactly 100 balls.

Urn A contains the exact same number of Red and Black balls (50 each).

The distribution of Red and Black balls in Urn B is unknown.

To specify your preferences between Urn A and Urn B, you only need to make one choice, namely in which row you would like to switch **from Urn A to Urn B**.

For each row **before** the one you have selected, a ball is drawn from Urn A. For each row **after** the one you have selected, including the one you have selected, Urn B is used for the draw.

Exactly one ball will be drawn. You will receive...

	Urn A **50 Red balls, 50 Black balls**	**Urn B** **? Red balls, ? Black balls**
1)	€20.00 if Chosen Color €0 if not	€16.40 if Chosen Color €0 if not
2)	€20.00 if Chosen Color €0 if not	€17.20 if Chosen Color €0 if not
3)	€20.00 if Chosen Color €0 if not	€18.00 if Chosen Color €0 if not
4)	€20.00 if Chosen Color €0 if not	€18.80 if Chosen Color €0 if not
5)	€20.00 if Chosen Color €0 if not	€19.60 if Chosen Color €0 if not
6)	€20.00 if Chosen Color €0 if not	€20.40 if Chosen Color €0 if not
7)	€20.00 if Chosen Color €0 if not	€21.20 if Chosen Color €0 if not
8)	€20.00 if Chosen Color €0 if not	€22.00 if Chosen Color €0 if not
9)	€20.00 if Chosen Color €0 if not	€22.80 if Chosen Color €0 if not
10)	€20.00 if Chosen Color €0 if not	€23.60 if Chosen Color €0 if not
11)	€20.00 if Chosen Color €0 if not	€24.40 if Chosen Color €0 if not
12)	€20.00 if Chosen Color €0 if not	€25.20 if Chosen Color €0 if not
13)	€20.00 if Chosen Color €0 if not	€26.00 if Chosen Color €0 if not
14)	€20.00 if Chosen Color €0 if not	€26.80 if Chosen Color €0 if not
15)	€20.00 if Chosen Color €0 if not	€27.60 if Chosen Color €0 if not
16)	€20.00 if Chosen Color €0 if not	€28.40 if Chosen Color €0 if not
17)	€20.00 if Chosen Color €0 if not	€29.20 if Chosen Color €0 if not
18)	€20.00 if Chosen Color €0 if not	€30.00 if Chosen Color €0 if not
19)	€20.00 if Chosen Color €0 if not	€30.80 if Chosen Color €0 if not
20)	€20.00 if Chosen Color €0 if not	€31.60 if Chosen Color €0 if not

Please indicate the decision for which you want to switch from Urn A to Urn B: ___

(The "Next" button appears after 60 seconds.)

***

Please answer the following questions.

What is your age?

What is your gender?

- ○ Male
- ○ Female
- ○ Non-binary

Please indicate the highest educational qualification you have obtained to date:

- ○ Middle school
- ○ High school/A-levels
- ○ Vocational training
- ○ Undergraduate degree
- ○ Graduate degree
- ○ Ph.D. or higher

Are you currently a university student?

- ○ Yes
- ○ No

What is your current study major or profession?

Is German your first language?

- ○ Yes
- ○ No

In general, how willing are you to take risks?

Not at all willing to take risks ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ Very willing to take risks

Is there anything else you would like to tell us? (optional)

***

Thank you for your participation!

Almost done...

For the evaluation of the experiment, it is important that we can rely on you having completed the experiment attentively and honestly.

If you did not complete the experiment attentively and honestly or if you completed it with interruptions, please let us know in the following questions.

Don't worry, you will still receive your reward!

Did you work through the experiment carefully?

- ○ Yes, I worked on the experiment carefully.
- ○ No, I did NOT work on the experiment carefully.

Did you complete the experiment in one go, i.e. without interruptions?

- ○ Yes, I completed the experiment in one go.
- ○ No, I did NOT complete the experiment in one go.

***

You have completed the experiment!

We will announce the winners of the bonus payment, as well as which payment mechanism was chosen for this, by circular mail to your Prolific accounts as soon as the number of participants reaches 100.

We will also inform you of your ranking in the test (only your Prolific ID will be displayed).

Thank you for your participation!

***