Dissociated Responses to AI: Legal AI Advisors Are Persuasive But Not Trustworthy?

By

Zeynep Aydin

B.S., Brown University, 2023

Thesis

Submitted in partial fulfillment of the requirements for the Degree of Master of Science in the Department of Cognitive, Linguistic, and Psychological Sciences at Brown University

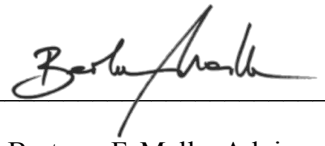
PROVIDENCE, RHODE ISLAND

MAY 2024

This thesis by Zeynep Aydin is accepted in its present form

by the Department of Cognitive, Linguistic, and Psychological Sciences as satisfying the

thesis requirement for the degree of Master of Science

Date May 1, 2024 

Dr. Bertram F. Malle, Advisor

Approved by the Graduate Council

Date_____ _____

Dr. Thomas A. Lewis, Director of Graduate Studies

Preface and Acknowledgments

I would like to take this opportunity to thank my advisor, Professor Bertram F. Malle, for his continuous support and guidance throughout my undergraduate and graduate research journey. He has not only been my thesis advisor but a valuable mentor. His wisdom, patience, and encouragement fostered my academic growth and confidence, shaping me not only as a scientist but also as a person. Thank you for inspiring me with your passion, attention to detail, and dedication; you have left an indelible mark on my career and life.

I would also like to express my gratitude to the readers of my thesis, Dr. Julia Netter and Dr. Steven Sloman, for their invaluable insights, broadening the scope of my research and thesis. I appreciate their profound questions and consistent availability.

I want to thank Dr. Joachim Krueger (Hocam), my freshman advisor and mentor, for guiding me throughout my academic career. Your insights into life, shared through Hoca Camide's wisdom, have consistently inspired me to question things and provided me with a broader perspective on life. Thank you for encouraging me to pursue my goals with confidence.

I owe a special thanks to Dr. Elena Festa, whose passion and care in her introductory class got me interested in psychology and cognitive science as an undergraduate. Thank you for introducing me to research and for your continued support.

I would like to thank all members of the Department of Cognitive Linguistic & Psychological Sciences and Computer Science for supporting me throughout my education and research journey.

Finally, my deepest thanks go to my dear parents and grandma. Thank you for believing in me, for your endless support, and for making countless sacrifices. I am grateful for how you nurtured my curiosity and fostered my love for learning from a young age. Thank you for your love and care. I love you all.

Table of Contents

Figures and Tables

The growing integration of artificial intelligence (AI) into society raises questions about its role: Should it be tools, assistants, or partners? Understanding people's psychological responses to AI agents is key to this debate. It is crucial for designing socially acceptable AI agents and sheds light on fundamental psychological processes, including social cognition, moral psychology, and social influence.

There is great interest in the research of autonomous decision-making capabilities of AI, which is especially active in healthcare and finance. For example, AI has proven effective in healthcare tasks like image-based cancer detection (Avendaño et al., 2023; Jones et al., 2022), disease prediction (Tasin et al., 2023; Zeng et al., 2022), and drug development (Soni & Hasija, 2022). In finance, AI applications include automated credit scoring (Gsenger & Strle, 2021), fraud detection (Cecchini et al., 2010), and investment decisions (Agrawal et al., 2019). Particularly in these fields, but also in many others, people encounter complex and morally significant decisions. AI proves useful by facilitating such decision problems that involve hundreds of decision factors with exponential combinations (Xu et al., 2020). Therefore, research has concentrated on developing formal theories and technologies to improve decision-making by adapting human capabilities for machine use (Xu et al., 2020) alongside exploring perceptions of AI's moral decision-making (Malle et al., 2015).

Alongside studies on AI decision-makers, artificial advisors are becoming increasingly popular (Hanson et al., 2024; Straßmann et al., 2020), signaling a potentially more imminent role for AI (Belazoui et al., 2022; Goel et al., 2023; Hwang et al., 2020; Kim et al., 2023). However, understanding how people use and rely on algorithms and artificial agents is crucial for developing socially acceptable and beneficial technologies. Consequently, this paper will focus on examining people's perceptions of AI advisors. I will delve into both explicit and implicit reactions to AI advisors, along with the factors that mediate these reactions. The initial chapters will provide an overview of AI advisors, exploring public reactions and perceptions, thus setting the stage for subsequent studies and analysis.

Chapter 1: AI Advisors

In many aspects of life, people rely on the help of advisors to make better-informed decisions, such as those concerning medical treatments or financial investments. As Artificial Intelligence becomes more prevalent, the role of AI advisors is expanding, either complementing or replacing traditional human experts or social circles. Such an integration of human insight and AI's computational capabilities may enhance decision outcomes. This collaborative approach, known as AI-assisted decision-making, leverages the unique strengths of both humans and AI agents (Zhang et al., 2020).

A notable field for AI-assisted decision-making is healthcare, where advisors can offer consistent guidance (Kim et al., 2023), targeted interventions (Belazoui et al., 2022), and help with the interactive collection of health information (Hwang et al., 2020). Beyond healthcare, AI advisors are advancing in various sectors. In finance, they assist in financial forecasting and provide insights for investment decisions, including portfolio optimization (Day et al., 2018; Goel et al., 2023). Similarly, in the legal sector, AI assists in court decisions, notoriously exemplified by the COMPAS criminal risk algorithm, which has faced criticism for its inaccuracies and biases (Dressel & Farid, 2018). These examples highlight that while AI applications are promising, they are not always accurate or beneficial. Therefore, it is crucial that individuals learn to appropriately rely on artificial advisors and discern between correct and incorrect guidance (Schemmer et al., 2022).

However, research has shown that human perceptions of AI are influenced by many factors that involve more than merely accepting or rejecting suggestions. People may react differently to the same advice from different advisors. For example, in moral dilemma scenarios like the Trolley Problem, robots are often blamed more for inaction than humans making the same decision (Hanson et al., 2024). This discrepancy underscores how the interplay between the source and content of advice influences judgments, which is essential for effectively developing and utilizing AI-driven systems (Bailey et al.,

2023; Schilbach et al., 2013). Empirical research came up with conflicting explanations on how algorithmic sources affect decisions. They may lead to an over-reliance due to their perceived objectivity, often resulting in "algorithm aversion." Conversely, there might be a reluctance to use algorithmic support even in cases where algorithms might outperform humans, a phenomenon often labeled as "algorithm appreciation." In the following section, I will delve deeper into the conditions that lead to the rise of either phenomenon and underscore the multifaceted nature of human responses to AI guidance.

Algorithm Aversion versus Appreciation

Algorithm aversion describes people's hesitance to trust or rely on AI for advice (Dietvorst et al., 2014; Jones-Jang & Park, 2023). This is especially evident in areas traditionally dominated by human judgment. For example, in medical settings, patients often prefer the recommendations of a human physician over those from a computer program, as people trust the physician to give better and more personalized advice (Promberger & Baron, 2006).

Several factors contribute to this phenomenon. People perceive artificial advisors as less capable of recognizing individual traits and circumstances than humans. This aversion tends to diminish when AI is framed as offering personalized care or when the AI is advising other patients rather than oneself (Longoni et al., 2019). This aligns with previous aversion findings suggesting that algorithms are perceived as dehumanizing or incapable of comprehending qualitative data (Grove & Meehl, 1996). Additionally, people tend to judge AI more harshly for mistakes compared to humans making the same errors (Renier et al., 2021). This discrepancy stems from the rapid decline in trust when people observe AI making errors. For example, trust in AI forecasters decreases more quickly than human forecasters after mistakes, even though AI may show superior forecasting accuracy overall (Dietvorst et al., 2015). Furthermore, people might have more general concerns about adopting AI technologies. Studies show that

people fear unemployment and are not comfortable with computers outperforming humans, influenced by their limited understanding of how algorithms work (Grove & Meehl, 1996).

In contrast, recent research suggests the emergence of an opposing trend known as algorithm appreciation, wherein individuals exhibit a greater reliance on AI compared to human advice (Hou & Jung, 2021; Logg et al., 2019; Schecter et al., 2023; Thurman et al., 2019). This concept was first introduced by Logg et al. (2019) when they demonstrated that individuals were more inclined to follow algorithmic advice across various subjective and objective tasks. For instance, participants preferred algorithmic guidance in tasks ranging from visual weight estimation to forecasting the popularity of songs, illustrating the breadth of algorithm appreciation across both objective and subjective domains (Logg et al., 2019). People's reliance on algorithms during estimation tasks further increases when prediction performance is present with minimal cognitive load, even when human and artificial advisors offer identical advice (You et al., 2022).

Overall, these two lines of findings, algorithm aversion and algorithm appreciation, suggest divergent systematic behaviors when individuals receive advice from AI versus humans. Therefore, their recommendations on how to best incorporate artificial agents differ. For example, increasing transparency about limitations of AI could be used to combat blind trust, whereas transparency on how AI generates personalized advice could mitigate aversion. Thus, I will introduce factors that underlie these phenomena in the following section.

The Emergence of Algorithm Aversion and Appreciation

Numerous factors influence the emergence of algorithm aversion or appreciation. Even in nearly identical decision-making scenarios, altering how human and AI advisors are portrayed can generate results that support either phenomenon (Hou & Jung, 2021). In this study, the researchers elicited algorithm

appreciation by presenting the artificial advisor as more competent than the human advisor. Conversely, shifting expert power to human advisors through subtle changes in their descriptions led to algorithm aversion. The key was the *relative* expert power attributed to each advisor, determining whether algorithm aversion or appreciation occurred. Therefore, presenting these agents as holding some expert power is crucial to fostering trust in artificial agents. This portrayal doesn't have to be explicit or verbal to shift people's perceptions. For instance, the credibility of Apple's Siri is perceived as low due to frequent errors in everyday interactions, which impacts how users respond to its advice. Even when it was presenting high-credibility information about the COVID-19 vaccine, using Siri's low-credibility voice in an experiment lead to reduced adherence to the advice (Dai et al., 2023).

Another crucial aspect to consider is the nature and architecture of algorithms. The lack of transparency surrounding machine learning, often described as the "black box" design, significantly contributes to aversion (Kayande et al., 2009). For example, when developers explain the workings of an algorithm and demonstrate how the system aligns with a shared vision and purpose, it can enhance trust in the artificial agent (Goodwin et al., 2013). However, the framing and extent of this explanation are crucial in alleviating aversion. While straightforward explanations can foster trust, complex descriptions of algorithms can still evoke unease (Stein et al., 2020). Additionally, there is a misconception that machines are incapable of learning, further exacerbating algorithm aversion (Dietvorst et al., 2015). However, when developers demonstrate the learning capabilities of an artificial system, it can enhance trust and reliance (Berger et al., 2021).

The anthropomorphism of artificial agents significantly influences users' perceptions. Agents that exhibit human-like characteristics tend to receive greater dependence and trust (Pak et al., 2012). This anthropomorphism extends beyond mere visual resemblance to humans, incorporating elements such as sound, language, and perceived liveliness (Pak et al., 2012; Stein et al., 2020). For example, synthesized voices are frequently perceived as less trustworthy and less likable (Dai et al., 2023; Kühne et al., 2020).

Moreover, the anthropomorphic qualities of artificial agents enhance users' perceptions of social presence, enhancing trust and enjoyment, thereby increasing their intention to use the agent as an advisor (Qiu & Benbasat, 2009).

Finally, in addition to the characteristics of the artificial agent, the type of task also plays a crucial role in the emergence of algorithm aversion or appreciation. Research shows that subjective tasks, such as evaluating art or making hiring decisions, often reduce reliance on algorithms because individuals mistakenly believe that algorithms are incapable of handling these nuanced tasks (Bower & Steyvers, 2021; Castelo et al., 2019). Conversely, in utilitarian tasks, where the goals are clear and quantifiable, like calculating loan eligibility or optimizing supply chain logistics, AI recommenders are perceived as more competent than human recommenders (Longoni & Cian, 2022). However, these task-dependent reactions can be moderated by changes in factors mentioned above, like anthropomorphism (Castelo et al., 2019). Moreover, recent studies indicate that individuals' processing of AI-generated information may not align with their affective responses—such as liking or trust—towards the AI (Bower & Steyvers, 2021; Liu & Moore, 2022; Renier et al., 2021). This dissociation adds complexity to the interpretation of findings related to aversion and appreciation, suggesting that developers must consider the interactions of multiple factors and varying responses to develop agents that generate an appropriate amount of trust and reliance.

The Role of Dual-Path Frameworks in AI Persuasion Research

The persuasion literature has long been interested in the extent to which individuals rely on heuristic cues versus engaging in more effortful processing of message content. Researchers have aimed to understand how these features interact to create persuasive influence. For example, involvement—which occurs when topics are personally significant or when recipients feel their opinions have important consequences—impacts persuasion outcomes, though with conflicting results. Chaiken (1980) observed that involvement could either enhance or diminish attitude changes, depending on heuristic cues, source

expertise, and the content of the message. Such complexities have led to the development of dual-path frameworks such as the heuristic-systematic model (Chaiken et al., 1989) and the elaboration-likelihood model (ELM; Petty & Cacioppo, 1986). These models provide insights into how people process persuasive messages, distinguishing between central (deliberative consideration of content) and peripheral (relying on affect and heuristics) processing routes. I am particularly inspired by these dual-path frameworks because they might offer a structured approach to explaining the conflicting results in the literature on algorithm aversion and appreciation.

While my studies don't use the specific tasks of the persuasion literature, I am using them as a guide to interpret my findings of dissociated responses to AI advisors. Various factors previously discussed—such as perceived advisor expertise—might trigger different processing paths depending on their emphasis. When expertise is explicitly mentioned, for example, it serve either as a central cue that prompts individuals to engage in effortful processing of the message content; however, if left implicit, it can also serve as a peripheral cue that influences the general liking or distrust of the information source. In my studies, I interpret persuasive message effects as indicating central information processing and AI versus human advisor source effects as indicating the peripheral processing route. Dual-path frameworks offer valuable terminology for exploring how responses to advice may vary between human and AI advisors.

Recent studies support this dynamic, showing that individuals adopt distinct strategies when interacting with AI-generated advice. Certain properties of the advisor, their tasks, and environments might lead to central processing, encouraging individuals to thoroughly evaluate the content and implications of AI advice, while in other contexts, they might depend on peripheral cues, such as the AI's anthropomorphic features or the task nature. The central-peripheral distinction helps explicitly account for these variations, explaining how specific attributes of AI can lead to different changes in behavior and trust levels. Here, it is also crucial to distinguish between reliance (shifts in behavior) and trust, as while they can be closely related, they are driven by different factors. Trust is mainly a subjective construct, developed from

expectations of the agent's capabilities, reliability, sincerity, and integrity (Malle & Ullman, 2021; Mayer et al., 1995), whereas reliance pertains to the decision to act on the advice, influenced by trust, but also convenience, necessity, etc. In the following chapter and the discussion of Study 1, I explore how dissociated paths and trust measurements inform my research in greater detail.

Chapter 2: Scope of Studies

In my research, I explored how people respond to legal advisors by applying a framework that distinguishes between dissociated paths of central information processing (effects from persuasive messages) and peripheral processing (differences in responses to AI versus human advisors) routes. In this chapter, I will discuss the role of AI advisors in the legal domain, emphasizing their promising potential. Subsequently, I will provide an overview of my studies.

AI Advisors in the Legal Domain

While much of the existing research on AI has focused on event forecasting (Logg et al., 2019; Önkal et al., 2009)l and medical diagnostics (e.g., Longoni et al., 2019; Promberger & Baron, 2006), only a limited number of studies have examined its use in legal proceedings, despite increasing real-world applications of AI in law (Al-Alawi & A-Lmansouri, 2023; Angwin et al., 2016; Roberts, 2023; Wang, 2020). AI holds promising potential in the courtroom, ranging from predicting trial outcomes to mitigating language barriers and predicting recidivism. Adapting AI to various legal cultures and systems may pose challenges, yet a well-implemented AI judiciary system could reduce biases, make proceedings more efficient, and lead to more informed decisions, benefiting all parties involved (Al-Alawi & A-Lmansouri, 2023). While people might be open to using algorithmic legal support, they remain averse to completely surrendering their decision-making authority to machines (Chugunova & Sele, 2022). Therefore, it is important to explore the various roles AI could play as an advisor in legal proceedings and the different

responses to it to ensure that the use of this technology yields favorable outcomes. While the reality of AI implementation in the legal field is still in its infancy, my research explores hypothetical, optimal scenarios involving AI advisors equipped with extensive knowledge bases and natural language capabilities to inform the development of technologies.

There is an intrinsic link between moral judgment and legal decision-making, as individuals' lives and social standings are at stake. Even though morally competent artificial agents are not yet fully implemented, empirical insights show differences in moral responses to AI and human advisors (Malle et al., 2019). They suggest that while people adjust their moral responses to human agents based on circumstances and justifications, their rationales for AI behavior are different, often because they do not readily see artificial agents as part of social structures. However, when individuals are reminded of the social contexts in which AI operates, they are more likely to modify their moral judgments accordingly (Malle et al., 2019). Understanding how individuals respond to AI advisors in morally charged legal contexts is crucial, especially as allowing individuals to choose which norms to apply to AI advice could lead to unpredictable judgment asymmetries, potentially causing misunderstandings or misplaced trust. Effective integration of (legal) AI advisors requires that individuals make informed and appropriate evaluations of the advice they receive. Therefore, my research draws from the literature on the moral competence of AI and employs a trust measure capable of discerning between performance and moral trust in Studies 2 and 3.

Design of the studies

Through a series of three studies, my thesis investigates people's responses to legal advisors—both human and artificial. The research framework employs a dual-path approach, which distinguishes between central information processing and peripheral processing, to assess how individuals form judgments and make decisions in legal AI advising contexts. Each study presents participants with carefully crafted legal

dilemmas, where two conflicting decisions (e.g., granting parole or not) are equally plausible, and participants provide their baseline support ratings (i.e., their initial stance). Participants then encounter arguments from either a human or AI legal advisor, favoring one decision or the other, and update their support. The changes from baseline to updated support measured how much participants shifted in the direction of the presented argument, constituting a "persuasion effect" (cf. Önkal et al., 2009; Prahl & Swol, 2021; Sniezek & Buckley, 1995, for similar paradigms). By measuring changes in participants' support for each decision, as well as their affective responses in terms of approval and perceived trustworthiness, the research explores potential differences in how individuals process information (the argument) and react affectively (approval and trust) when influenced by human versus AI advisors.

The studies use various scenarios and argument presentations to examine potential aversion, appreciation, and dissociated response routes toward AI advisors. Pretests ensured the credibility of the legal dilemmas presented, with a careful selection of cases featuring equal support for each decision. Designs differ slightly across the three studies. Study 1 presented participants with two legal cases featuring the same type of advisor (human or AI). Study 2 focused on one case with opposing arguments from human and AI advisors and trust ratings, and Study 3 explored the effects of varying argument strengths within a single case. This comprehensive approach is helpful in developing an understanding of how individuals react to AI advisors in a legal decision-making context, ranging from simple responses to more complex interactions.

Chapter 3: Study 1 - Responses to AI Advisors

Methods

Participants

In total, I recruited 217 participants through the online crowdsourcing platform Prolific. The objective was to secure a minimum of 100 participants for each condition—human or AI advisor—with an anticipated effect size of d = 0.40 or higher. After excluding 20 cases that didn't pass the bot check, the analysis included 197 valid participants. Of these, 90 identified as female and 101 as male, with an average age of 41 years (SD = 14.5). Each participant received $1.00 as compensation for completing the 6-minute survey.

Stimulus Selection

I conducted a pilot study to identify the most suitable stimuli. Initially, I crafted six candidate cases based on real legal cases (e.g., *Clark v. Arizona*, 2006; *People v. Barnes*, 1990; *People v. Watson*, 1981), each presenting a challenging decision (e.g., granting parole or not, lenient vs. harsh sentence) where could be reasonably defended. These cases were broadly categorized into parole, court process, and conviction, with two instances for each.

During the pilot study, participants encountered three scenarios (one at a time), with each scenario randomly selected from one of the three categories and the order of categories randomized. Following the reading of a scenario, participants expressed their level of support for one decision or the other on a 100-point scale, with the poles representing opposing decisions.

I enlisted 162 participants via Prolific for the pretest sample. The results showed that two of the six cases were promising, as the average support ratings were close to the midpoint of 50. The selected scenarios

included a parole case (M = 55.4) and an insanity plea case (M = 49.9). Conversely, the remaining scenarios were excluded due to their support rates averaging over 70.0 across all options and were not included in the final study.

Procedure and Measures

After the consent procedure, participants are randomly assigned to either the AI or human legal advisor condition. They received a brief introduction tailored to their assigned condition, outlining the use of legal advisors, whether AI or human, in legal proceedings as follows:

> "*The use of* [Artificial Intelligence (AI) / legal] *as advisors in legal proceedings is an intriguing possibility. A number of people argue that the* [use of AI / it] *will make legal decisions less biased, whereas others argue that it will make them more biased.*
> *We present you with two challenging decision scenarios (shortened from real cases) in which* [an AI / a legal] *advisor proposes a certain decision. After each case, we will ask you about your impressions."*

Next, participants read one Case (either parole or insanity plea, counterbalanced) and expressed their initial support for the decision (baseline support) using a slider bar. For example, they answered the question, "How strongly do you believe that Richard K. should or should not be granted parole?" on a 0-100 slider scale, with 0 labeled "Definitely not grant parole" and 100 labeled "Definitely grant parole."

Following this, participants received a recommendation from their assigned Advisor (AI or human), endorsing a randomly assigned Argument direction (e.g., for or against parole). Subsequently, participants indicated their level of support for either direction, now integrating the advisor's recommendation (updated support), and expressed their approval of the advisor ("To what extent do you approve of the AI advisor giving this specific advice?"), both responses were measured on a scale of 0 to 100. An example illustration of the study's progression up to this point is shown in *Fig. 1*, with between-subject agent manipulations separated by square brackets.

Next, participants received the other case, where the type of advisor stayed the same, but the argument direction was opposite to that of the first case. Participants completed identical assessments of baseline and updated support alongside advisor approval for this second scenario. Finally, participants provided demographic information about age, gender, AI knowledge, and programming experience.

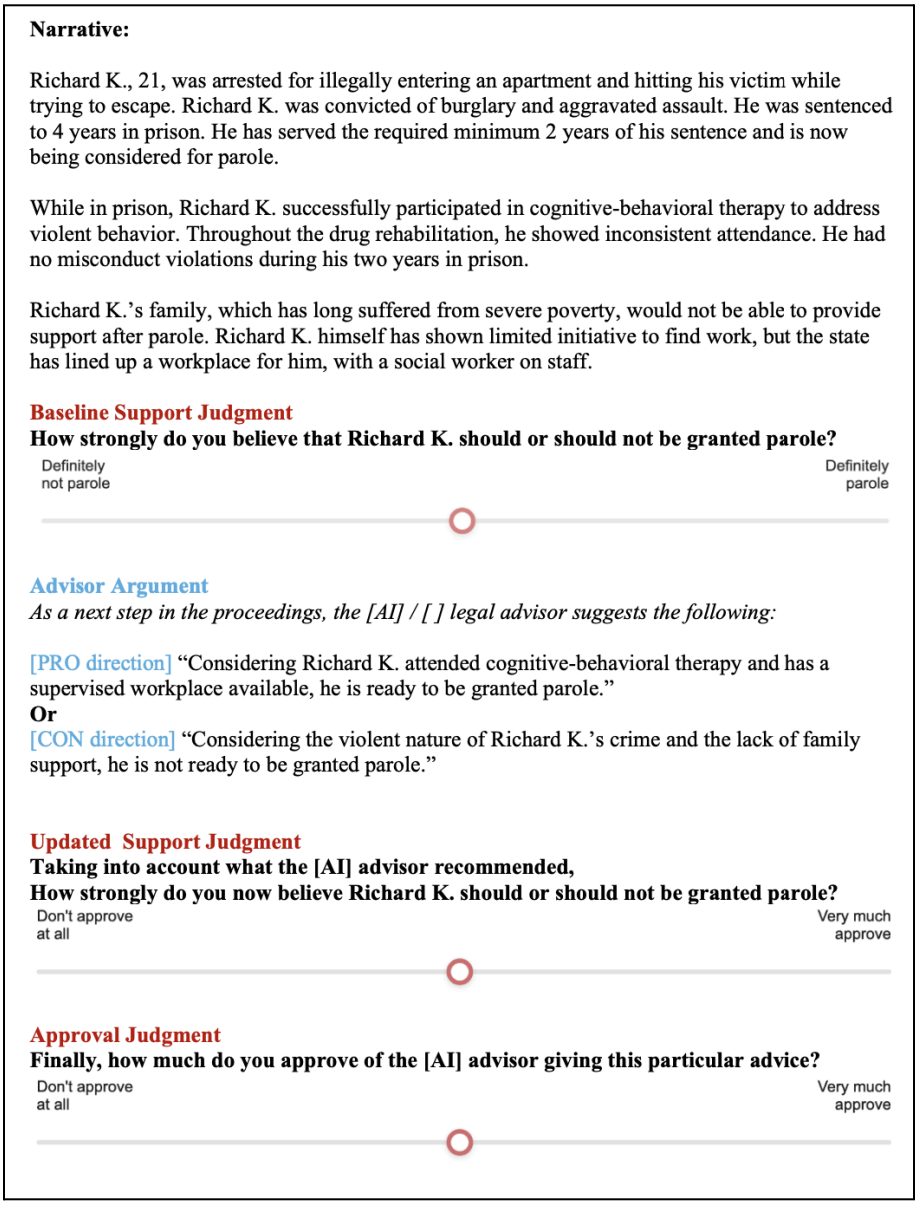


Fig. 1: *Experimental material (narrative, arguments, and dependent variables) for Study 1. The between-subjects manipulation of Advisor (AI advisor or human legal advisor) is indicated by different square brackets; the between-subjects manipulation of Direction (launch the strike vs. cancel the strike) is indicated by blue square brackets. The within-subject measures are marked with red titles.*

Results

I conducted separate mixed-design ANOVAs for each legal case, incorporating the between-subjects factors of Advisor (AI vs. human) and Argument (pro vs. con) and the within-subject factor of Support change (baseline to updated). For simplicity, I report analyses on the support change scores (which are mathematically identical to the full mixed-design results).

Participants were slightly inclined towards granting parole (M = 54.6, SD = 27.8), deviating from the neutral point of 50, $t(195) = 2.30$, $p = 0.023$. Conversely, they leaned slightly against accepting the insanity plea (M = 44.4, SD = 29.8), $t(195) = -2.60$, $p = 0.01$. Support ratings for the two cases were largely independent, $r = 0.167$ ($p = 0.019$).

Persuasion Effects

The persuasion effect in my study aims to measure how much participants shifted their support toward the presented arguments—either pro or con. To achieve this, I initially collected participants' responses to the scenario (baseline support scores). Subsequently, I asked for their updated support ratings after they received the advice using the same scale. The difference between these two scores constituted the persuasion effect. I conducted separate ANOVAs of these change scores for each legal case, predicted by the between-subjects factors of Advisor (AI vs. human) and Argument (pro vs. con).

In the case of the parole scenario, the support change revealed a persuasion effect of the pro-argument, increasing support (+5.1 points on a 0-100 scale), and the con-argument, decreasing support (-8.9 points), $F(1, 192) = 65.4$, $p < .001$, $\eta^2 = 25.4\%$. No significant interaction was observed with the type of Advisor, $F < 1$, $\eta^2 < 0.1\%$, as the persuasion effect was highly similar for both the AI advisor (d = 1.24) and the human advisor (d = 1.07). A follow-up Bayesian analysis confirmed that there was strong evidence against an interaction effect (BF = 0.251).

The insanity plea case similarly demonstrated a persuasion effect, with pro-arguments increasing support (+8.8) and con-arguments decreasing support (-7.3), $F(1, 192) = 47.9$, $p < .001$, $\eta^2 = 20\%$. There was no significant interaction with the type of advisor, $F(1, 192) = 1.9$, $p = 0.17$, $\eta^2 = 1\%$. However, in a numerical comparison, the persuasion effect tended to be stronger for the human advisor ($d = 1.19$) than for the AI advisor ($d = 0.79$). A follow-up Bayesian analysis provided no conclusive evidence against an interaction effect ($BF = 0.513$).

The similarity of support change between the cases is shown in *Fig 2*.

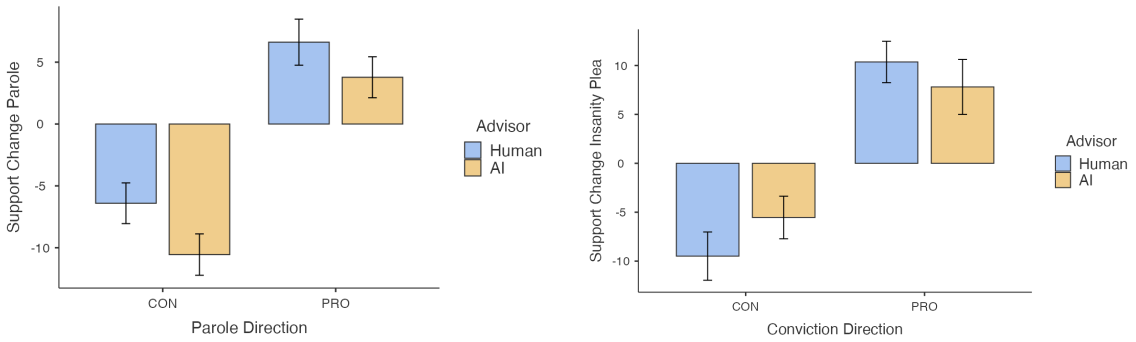


Fig 2: *Study 1 shows similar persuasion effects for AI and Human advisors in the parole case (left) and the insanity plea case (right). Error bars are 95% CIs.*

Approval

I measured Approval by directly asking participants how much they approved of the advisor after they provided their updated judgment ratings. While the persuasion effects were equally strong for both AI and human advisors, participants tended to express less approval toward the AI advisor compared to the human advisor (see Fig 3).

In the parole case, the AI advisor received an approval rating of $M = 54.1$, whereas the human advisor received a higher approval rating of $M = 62.0$, though this difference did not reach statistical significance,

$F(1, 192) = 3.48$, $p = 0.064$, $d = 0.27$. Similarly, in the insanity plea case, the AI advisor received lower approval ratings ($M = 49.1$) compared to the human advisor ($M = 59.0$), with the difference being statistically significant, $F(1, 192) = 5.47$, $p = 0.02$, $d = 0.33$. Neither of these approval differences interacted with the direction of the argument ($ps > .28$, η^2s < 1%), indicating that the variation in approval was attributable to the type of advisor rather than the argument presented by the advisor.

Approval ratings and support change scores were uncorrelated, both for parole (full sample $r = -0.12$, Human $r = -0.17$, AI $r = -0.11$) and for insanity plea (full sample $r = .05$, Human r =0.02, AI $r = 0.12$), with all $ps > 0.11$.

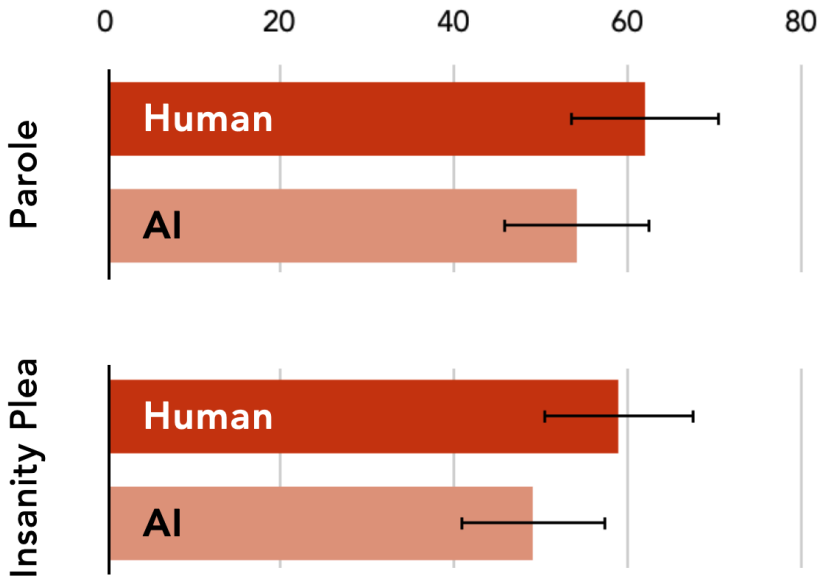


Fig 3: *Study 1 shows greater approval for the human advisor than for the AI advisor in the parole case (top) and the insanity plea case (bottom). Error bars are 95% CIs.*

Discussion

The two legal cases elicited the intended decision conflict, as predicted by the pilot studies. This was evident from the sample distribution of support, which was roughly centered on the midpoint of the scale. However, upon exposure to an advisor's argument advocating for one side or the other, people were

effectively persuaded, leading to a shift in their support toward the advocated position. Notably, this persuasion effect was equally strong for human and AI advisors, underscoring the primary influence of the advice itself rather than the source delivering it. This suggests that there is no discernible pattern of algorithm aversion or appreciation in scenarios involving central-route information processing, where individuals carefully consider message contents.

By contrast, approval ratings revealed a different pattern, where participants tended to express lower approval ratings for the AI advisor than the human advisor, regardless of the argument direction or the specific legal case. This discrepancy in approval ratings indicates that people had affective reservations of approximately 8-10 points towards the AI advisor along what we might presume to be a peripheral route." Furthermore, support change and approval ratings appeared to be uncorrelated, suggesting that the two routes of persuasion were independent of each other.

However, this study does not offer a direct comparison between human and AI advisors. Could the equally sized persuasion effect for both human and AI advisors have been facilitated by the fact that participants only considered one advisor (and one argument) at a time? To examine this possibility, Study 2 invited people to directly compare the two advisors giving opposing arguments. This setup might expose affective reservations towards AI, potentially distorting participants' processing of message content. Drawing from classic source effects in the persuasion literature, people might perceive an argument presented by a human as more persuasive than the opposing argument presented by AI, irrespective of the actual content or source (Dai et al., 2023; Feng & MacGeorge, 2010).

In Study 2, I aimed to address another critical aspect: establishing a metric for trust. In the literature, what I referred to as the "persuasion effect" is commonly known as "advice utilization" and is often considered a metric for trust in the advisor (Bonaccio & Dalal, 2006; Sniezek & Van Swol, 2001). For example, decision-makers are inclined to ignore advice if they have doubts about the advisor's trustworthiness, such

as suspecting misaligned goals (Bonaccio & Dalal, 2006). However, it is important to differentiate between trust and reliance. While trust reflects a subjective expectation of the advisor's capabilities, reliability, sincerity, and integrity (Malle & Ullman, 2021; Mayer et al., 1995), reliance pertains to the decision to act on the advice, influenced by factors like trust, convenience, or necessity.

Recent research has shown that trust, as a subjective expectation, is made up of multiple dimensions, including performance trust (the belief in an agent's competence and reliability) and moral trust (the belief in an agent's ethical, sincere, and benevolent nature). This multi-dimensional perspective applies to trust in both humans and machines. By measuring trust in this multi-dimensional manner, I gain insights into another aspect of AI aversion (or appreciation). Performance trust reflects the agent's capability in handling content-related tasks, the absence of which appears to be a significant factor in AI aversion (Hou & Jung, 2021; Parasuraman & Riley, 1997), while moral trust encompasses broader social-relational capabilities (Law & Scheutz, 2021; Malle & Ullman, 2021). Some studies suggest that people may particularly reject AI in moral domains (Bigman & Gray, 2018), and given that many legal issues involve moral considerations, it is relevant to explore this dimension. Therefore, in Study 2, I added a multi-dimensional trust measure and anticipated that trust, particularly moral trust, would mirror approval ratings, showing lower levels for AI compared to human advisors. This difference along the peripheral processing route should contrast with the persuasive effect observed equally for AI and human advisors along the central processing route.

Chapter 4: Study 2 - Dissociated Responses to AI Advisors

In this study, participants encountered a single scenario, but they received counsel from both agents, a human and an AI advisor, giving opposing arguments. I evaluated the degree to which participants were persuaded by one or the other advisor. Additionally, I examined approval ratings and trust measures. To ensure robustness, I counterbalanced other factors, such as pairing each advisor with a specific argument and varying the order in which advisors presented their arguments.

Methods

Participants

In total, I recruited 247 participants through the online crowdsourcing platform Prolific (participants of Study 1 were not eligible to participate in this study). After excluding 11 empty responses and 23 cases that didn't pass the bot check, the analysis included 211 valid participants. Of these, 126 identified as female and 83 as male, with an average age of 41.6 years (SD = 14.2). Each participant received $1.40 as compensation for completing the 7-minute survey.

Procedure and Measures

I maintained the same two legal dilemmas used in Study 1, and the measures and procedures remained similar. However, Study 2 introduced significant changes and additions. Participants were presented with only one of the cases (either parole or insanity plea), but they received two opposing arguments: one from a human legal advisor and the opposite one from an AI advisor. These arguments, identical to those used

in Study 1, have been pretested to ensure comparable persuasive strength, resulting in a strong advice opposition between human and AI advisors.

Once the participants completed the consent procedure, they received the following introduction:

> "*The use of Artificial Intelligence (AI) as advisors in legal proceedings alongside human advisors is an intriguing possibility.*
>
> *We present you with a challenging decision scenario (shortened from a real case) in which both human and AI advisors propose a certain decision. After the case, we will ask you about your impressions."*

Participants proceeded by reading one of the Cases (either parole or insanity plea, counterbalanced) and indicating their initial support for the decision using a slider bar, mirroring the procedure in Study 1. Then, they were presented with the argument from the first advisor (human or AI, counterbalanced), who advocated for one side of the case (randomly assigned to be pro vs. con). Following this first argument, participants updated their support for the decision and provided their approval ratings of the first advisor, consistent with the measures employed in Study 1. Next, they encountered the opposing advisor advocating for the alternative side. Once more, participants updated their support for the decision and expressed their approval of the second advisor. After this section, participants completed a trust measure for one of the advisors and then a parallel form of that trust measure for the other advisor (both order and forms were counterbalanced). A flowchart of the study's progression up to this point is shown in *Fig. 4*, which uses the insanity plea case as an example. Finally, participants provided demographic information and indicated their level of experience with AI and programming.

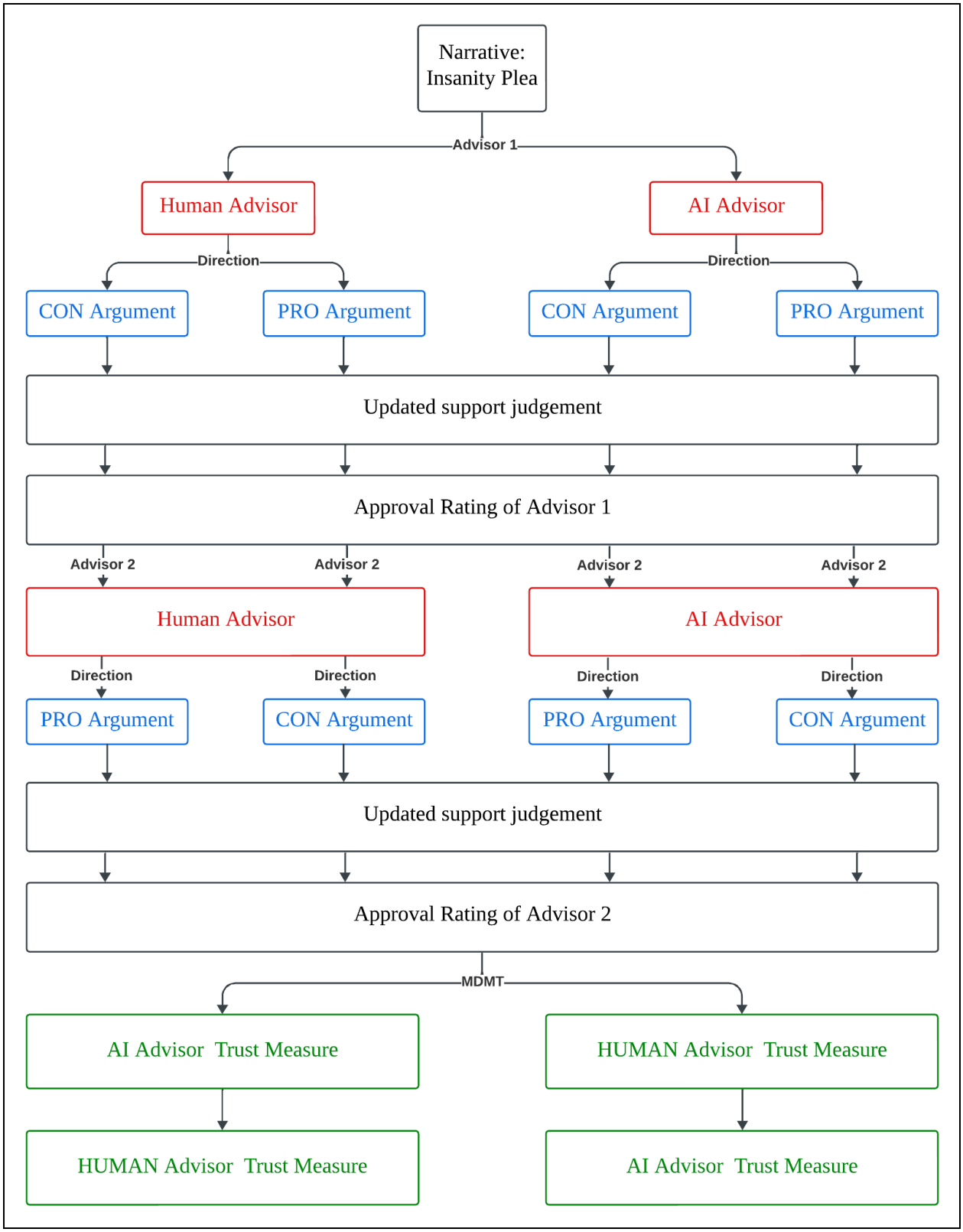


Fig. 4: *Flowchart for Study 2.*

In the Study 1 Discussion, I delved into the complex nature of trust, emphasizing the need for a more precise measurement approach beyond simple approval or support change inquiries. Therefore, to capture the essence of trust, I employed the Multi-Dimensional Measure of Trust (MDMT v2; Ullman & Malle, 2018, 2023), which conceptualizes trust as a multi-dimensional set of expectations regarding the agent's trustworthy dispositions (Malle & Ullman, 2021). This tool is specifically designed to assess trust in both human and artificial agents across five dimensions, organized into two major factors: Performance trust (encompassing Competence and Reliability) and Moral trust (covering Ethicality, Transparency, and Benevolence).

The MDMT v2 demonstrates strong internal consistency, which is evident at both the factor and dimension subscale levels. However, due to the high correlations among subscales, I focused my analysis on the Performance trust and Moral trust factors. Since participants completed the trust measure twice, once for each advisor, I used the two parallel 10-item short forms of the MDMT. These short forms feature entirely distinct items but maintain a high level of correlation.

Design and Analysis

Each participant encountered only one of the two Cases (parole vs. insanity plea) and provided baseline support for the case. After the two arguments, they reported updated support judgments on the case and approval ratings for both the human and AI advisors. Participants received counsel with one of the Argument compositions (pro for AI and con for human advisor or vice versa, with the order of advisors randomized). Thus, Case and Argument were between-subjects factors, whereas Advisor type was within-subjects. Trust measurements were collected using the MDMT, with the order of forms and advisors randomized. Overall, I counterbalanced the Argument order (human vs. AI offers argument first), argument direction (pro vs. con argument first), trust assessment order (human vs. AI trust assessed first), and trust form (MDMT A vs. B). For the primary analyses, the two cases were examined separately, using a mixed within-between ANOVA of Advisor × Argument composition × Argument order.

Results

At baseline, participants exhibited a slight inclination against granting parole ($M = 44.7$, $SD = 26.6$), diverging from the neutral 50, $t(104) = -2.04$, $p = 0.043$. Conversely, they were nearly neutral towards accepting or denying the insanity plea ($M = 47.0$, $SD = 28.8$), $t(105) = -1.09$, $p = 0.28$.

Persuasion Effects

I began by analyzing the parole case (Fig. 5, left). Overall, support for parole exhibited a slight downward trend of 1.74 points ($p = .064$). The anticipated persuasion effect emerged, indicating that individuals who received a pro-argument updated their support upwards for both the human advisor ($M = +2.39$) and the AI advisor ($M = +1.33$). Similarly, those who received a con-argument altered their support downwards for both human ($M = -5.67$) and AI advisors ($M = -5.02$), $F(1, 100) = 44.8$, $p < .001$, $\eta^2 = 30.9\%$. These patterns are technically interactions, given that the advisors' arguments are consistently opposite in the study design.

Furthermore, I saw a primacy effect, indicating that the first argument, regardless of its source, was consistently more persuasive ($F(1, 100) = 9.0$, $p = .003$, $\eta^2 = 8.2\%$). Specifically, after being presented with a pro-parole argument first, individuals increased their support to a greater extent ($M = +3.81$) than after being presented with it second ($M = -0.15$). Conversely, after being presented with the con-argument first, individuals decreased their support more substantially ($M = -8.71$) than after being presented with it second ($M = -0.96$).

In the insanity plea case (Fig. 6, right), there was a tendency for support for the plea to increase by 2.64 points from baseline ($p = 0.11$). Once again, a persuasion effect was evident: individuals who received the pro-argument demonstrated an upward change both for human ($M = +8.52$) and AI ($M = +6.05$) advisors. Likewise, those who received a con-argument underwent a downward change for both advisors, human

(M = -1.02) and AI (M = -4.03), resulting in a significant effect (F(1, 103) = 17.5, p < .001, η^2 = 14.5%). Notably, no primacy effect has emerged.

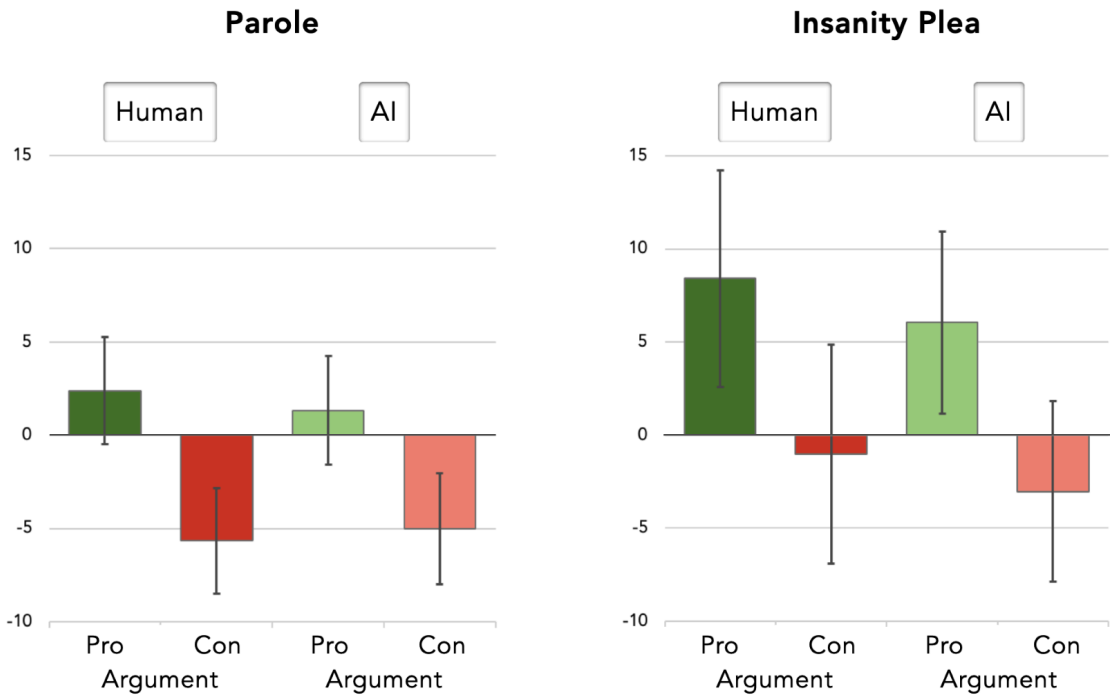


Fig 5: *Study 2 shows similar persuasion effects (support change from baseline) from AI and human advisors in the parole case (left) and the insanity plea case (right). Pro-arguments increase support, and con-arguments decrease support, no matter who advises. Error bars are 95% CIs.*

Approval

Similar to Study 1, approval judgments remained distinct from the persuasion effects (see Fig 6). In the parole case, individuals favored the human advisor more (M = 62.1) than the AI advisor (M = 50.6), revealing a significant difference (F(1, 100) = 5.65, p = .019, d = 0.41). Similarly, in the insanity plea case, participants showed greater approval for the human advisor (M = 68.7) compared to the AI advisor (M = 45.4), with a notable effect size (F(1, 102) = 24.0, p < .001, d = 0.82).

However, an interesting outcome in the parole case was that one condition resulted in nearly equal approval ratings for both human and AI advisors (see dotted markings in Fig 6, left panel). This occurred

because, in the parole case, the con-arguments received overall higher approval than the pro-arguments. Thus, in the condition where the AI's con-argument was contrasted with the human's pro-argument, the AI's overall approval deficit was offset by the lesser approval for the pro-argument. This exception to the general lower AI approval did not manifest in the insanity plea case since participants exhibited no preference for the pro- or con-arguments. The only consistent pattern was that the AI advisor received lower approval than the human advisor in both cases. Notably, there was no observed order effect concerning which advisor presented their argument first.

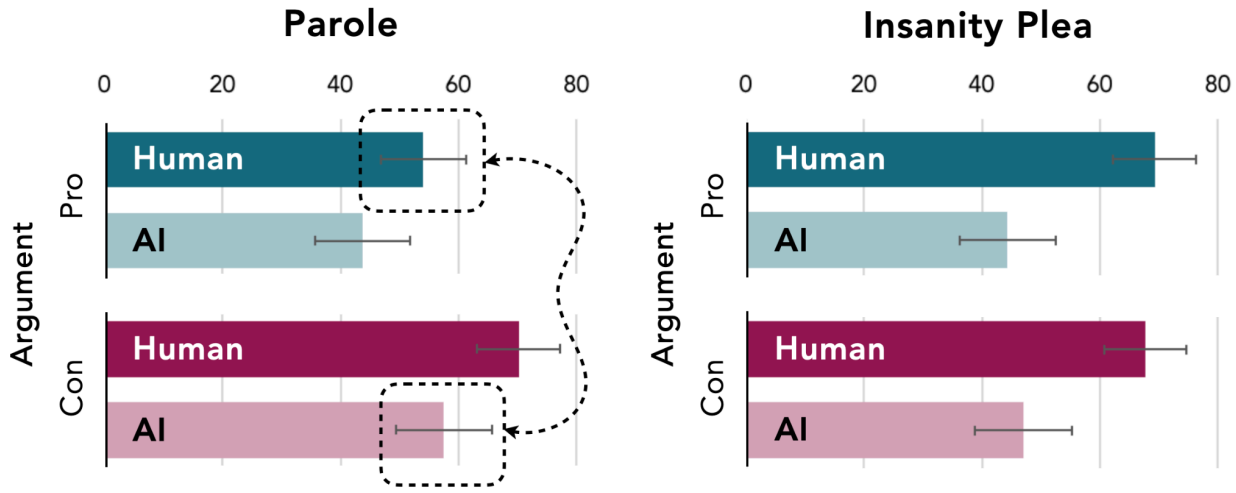


Fig 6: *Study 2 shows consistently higher approval for the human than the AI advisor for the same arguments (pro or con), both for the parole case (left) and the insanity plea case (right). The dotted lines mark a seeming exception (see text for interpretation). Error bars are 95% CIs.*

Trust

Participants completed their trust ratings after providing all support and approval judgments. I report the aggregated findings of the two cases since the analysis showed no noteworthy differences between them. The first finding revealed considerably higher overall trust in the human advisor compared to the AI advisor, $F(1, 195) = 46.1$, $p < .001$, $\eta^2 = 19.1\%$. This was evident in both slightly higher Performance trust in the human ($M = 3.45$) compared to the AI ($M = 3.12$), $d = 0.43$, and substantially higher Moral trust in the human ($M = 3.47$) compared to the AI ($M = 2.54$), $F(1, 195) = 44.7$, $p < .001$, $d = 0.81$.

As shown in Fig. 8, the deficit in trust toward the AI advisor was mitigated when the AI advisor presented con-arguments, $F(1, 195) = 7.2$, $p = .008$, with an interaction $\eta^2 = 3.5\%$. However, this mitigation effect was specific to Performance trust, $F(1, 195) = 18.3$, $p < .001$, $\eta^2 = 8.6\%$. Interestingly, the distrust toward the AI advisor was lessened by con-arguments across both legal cases, despite a slight preference for pro-arguments in the insanity plea case (see Fig. 6). This suggests that trust, similar to approval, is independent of argument content, as the AI is perceived as more competent when advocating against parole or against the insanity plea.

In additional analyses, I explored whether the order of presentation of the trust measure had an impact on the AI trust deficit. Indeed, the deficit in trust toward AI was weaker when AI trust ratings came first: $F(1, 195) = 6.8$, $p = .01$, $\eta^2 = 3.4\%$. This finding suggests that participants, when asked to express trust in AI first, might have emphasized the positive aspects of AI in contrast to what it lacks, particularly when compared to human advisors.

The differences in overall approval and trust trends underscore the importance of including the trust measure, as simple approval or support change does not directly reflect trust, a point also mentioned in the Study 1 discussion.

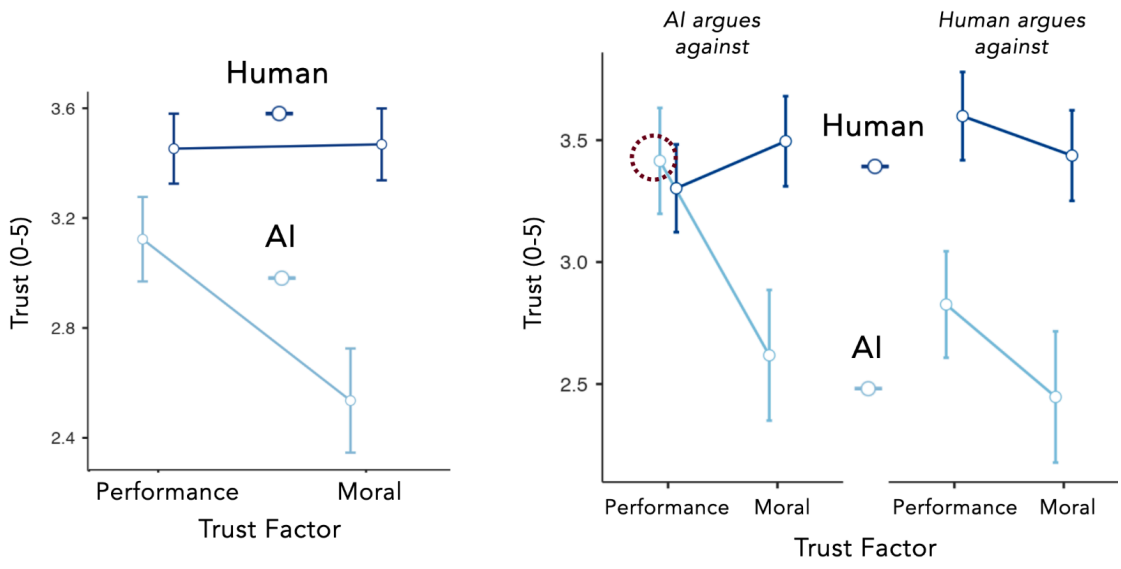


Fig 7: *The left panel shows higher trust ratings in the human advisor than the AI advisor in Study 2, especially for moral trust. The right panel shows mitigation of this AI trust deficit for performance trust when the AI argues against a decision (dotted circle), no matter which decision.*

Relationship Among Measures

Finally, I examined how the three main measures related to each other. Table 1 shows that the content-based support update is orthogonal to the more affective agent variables of approval and trust, a pattern observed for both human and AI advisors. As a result, controlling for the affective variables does not change the results of equal human-AI support updates.

However, there is a close relationship between approval and trust (*r*s between .46 and .56). Therefore, I examined whether the measures are redundant. The AI's lower approval decreases from an effect size of η^2 = 11.1% to 5.7% when controlling for the AI's Performance trust deficit, and it decreases to 0% when controlling for the AI's Moral trust deficit. Hence, the disapproval of AI is primarily rooted in perceived moral shortcomings rather than perceived competence limitations. The reverse analysis, controlling for approval in examining the AI trust deficits, shows that the overall trust deficit reduces from 19% to 10%. However, the greater Moral than Performance trust deficit remains unchanged, as does the mitigation of

performance trust when the AI offers a con-argument (Figure 7). Thus, approval judgments seem to reflect trust concerns more than trust concerns reflect approval judgments.

Table 1: *Correlations among main measures in Study 2*

	Support Update	Approval	Performance Trust	Moral Trust
Support Update		-0.06	-0.10	-0.05
Approval	0.00		**0.53**	**0.47**
Performance Trust	0.02	**0.56**		**0.60**
Moral Trust	0.03	**0.46**	**0.74**	

Note. Human values are in the lower triangle, and AI values are in the upper triangle. For boldfaced values, $p < .001$; for all others, $p > .05$.

Discussion

Study 2 directly pitted opposing human and AI advisor arguments of equal strength against each other. Using the same legal scenarios as in Study 1, I assessed content-based support judgments (measured as the difference in support before and after the argument), as well as approval and trust. Consistent with Study 1, arguments advocating for or against the legal decisions persuaded participants to shift their judgments, irrespective of whether they originated from a human or AI advisor. However, also akin to Study 1, participants consistently exhibited less approval for the AI advisor compared to the human advisor. Furthermore, trust revealed a parallel AI deficit, albeit to a lesser extent in Performance trust (pertaining to competence and reliability) than in Moral trust (reflecting integrity, transparency, and benevolence). Hence, the pattern of findings indicates a dissociation between content-based persuasion (equally strong regardless of the agent's identity) and agent-specific sentiments (showing greater approval and trust for human advisors over AI).

Under one specific condition (see Figure 8), the unfavorable sentiments toward the AI advisor disappeared: Performance trust in AI was equal to Performance trust in humans when the AI advocated

against granting parole or accepting the insanity plea—possibly because opposing such requests conveys a sense of confidence or expertise. Under one other condition, AI approval increased to the human level (see parole case in Figure 7), but this pattern is fully explained by the fact that the AI provided advice aligned with the slightly more popular decision, and the human advisor offered advice contrary to the popular decision. Thus, the (affective) aversion to the AI advisor was offset by the (cognitive) effect of the argument. In this particular context, the affective (peripheral) and cognitive (central) processing routes interacted, despite the overarching trend of dissociation between the two.

The studies so far have been using strong, well-constructed arguments. Could it be that the arguments are so strong that the source (AI or human) is irrelevant? Is it possible that the observed patterns would differ if one advisor (or both) presented a particularly weak argument? Moreover, the studies so far have been using arguments that are closely modeled after the legal case information presented at the outset of the study. Could the particular attributes of the argument, such as its strength or type, be the primary factor driving the equal response (support change) to the arguments, overshadowing the source effects? These were the primary questions motivating Study 3. In addition, I explored whether individuals hold specific expectations about the types of arguments they might receive from human versus AI advisors. For instance, if an argument aligns with the common perception of AI as being data- and numbers-driven, does this alignment increase trust? To explore these possibilities, Study 3 employed a purely between-subjects design, inviting participants to evaluate cases where they received either a well- or poorly-constructed argument of one of three possible types: specific to the case at hand, statistical, or referencing to another case.

Pitting a good argument against a bad argument also offers an opportunity to apply a different theoretical framework to analyze the observed dissociation: Salmon's (1971) concept of screening-off. This concept states that in a causal chain from a distal cause (D) to a proximal cause (P) to an effect (E), P is considered to "screen off" D from E when $Pr(E|P\&D) = Pr(E|P) \neq Pr(E|D)$ (Sober, 1992). In my case, the

causal chain includes the Advisor (Source, $S_{AI,}$ or S_H) as the distal cause, the Argument (Evidence, E_{Good} or E_{Bad}) as the proximal cause, and the Support for the case decision as the Effect. According to the screen-off principle, if the Argument screens off the Advisor, then both good and bad advice would lead to the same (differential) Support, regardless of Advisor type. If $Pr(A|E_{Good} \& S_H) = P(H|E_{Good} \& S_H)$ and $Pr(A|E_{Bad} \& S_H) = P(H|E_{Bad} \& S_H)$, it suggests that people are influenced solely by the content of the argument, disregarding the source. However, if I find no (or only a partial) screening-off effect, it indicates that both the Advisor and the Argument jointly influence the resulting Attitude.

Chapter 5: Study 3 - Ongoing

In this study, participants engaged in a single scenario where they received counsel from either a human or an AI advisor. Each advisor presented an argument advocating for either a pro or con decision, varying in quality (good or bad) and using different types of information (case-specific, statistical, or referencing another case). The study assessed the extent to which participants were persuaded by the advisor's argument. Additionally, the research examined participants' approval ratings and trust measures toward the advisor.

I anticipate that the quality of arguments will have an influence on participants' updated support for parole, independent of other factors. Drawing from the consistent findings of Studies 1 and 2, where participants' opinions were similarly influenced by both AI and human advisors, I predict that the impact of argument quality, irrespective of whether presented by an AI or human advisor, will remain consistent. Additionally, while considering the differential effects of argument type and direction on updated support, I do not expect these effects to be moderated by the agent factor (AI vs. human advisor). Furthermore, I anticipate observing lower levels of moral trust in AI advisors compared to human advisors, regardless of the strength, type, or direction of the argument. Lastly, I expect to find variations in performance trust based on the quality of the arguments presented.

The data collection for this study has not yet been completed. Therefore, I will only present the methods section below.

Participants

Participants will be recruited through the online crowdsourcing platform Prolific. Planned sample size calculations are based on the previous AI deficit in trust because the argument comparisons are new, but we wanted to ensure that the AI deficit, at least for Moral trust, could be replicated. In Study 2, the AI deficit was $d = 0.42$ for Performance trust and $d = 0.79$, so the objective is to secure a minimum of 62 participants in each of the human and AI conditions to detect between-subjects effect sizes of $d \geq .50$ at $p < .05$ (two-sided) and power $\geq .80$ in the cell that replicates Study 2 (good argument quality, current-case evidence). Since I have six cells total in the 2 (argument quality) x 3 (argument type) design, I will recruit 780 participants, rounding up to account for possible exclusions. For a given argument type, this sample provides power $\geq .80$ at $p < .05$ (two-sided) to detect argument quality differences (across agents) of $d \geq 0.35$.

Argument Selection

I crafted arguments of varying quality (good and bad) and type (specific to the case at hand, statistical, or referencing another case) that were either pro or con towards granting parole (the same scenario used in Studies 1 and 2).

In a pilot study, participants encountered the parole scenario and gave their baseline support judgments. Subsequently, they received one of the crafted pieces of advice from a (human) legal advisor, updated their judgments based on the advice they received, and provided ratings regarding the strength, influence,

and persuasiveness of the advice. For instance, they answered, “How strong is the advisor’s argument?” and “How much did the advisor’s argument influence your own opinion?” on a 100-point scale.

I recruited a sample of 287 participants through Prolific for the pretest phase. The findings revealed that argument quality indeed exerted the anticipated influence, with well-constructed arguments demonstrating a greater impact. Specifically, all comparisons of good versus bad quality for Con arguments yielded effect sizes of $d \geq 0.34$, but two Pro-argument comparisons (case-specific and reference) did not meet this threshold. Consequently, for the main study, I clarified the language and strengthened the good-quality arguments. Furthermore, the considerable overlap between responses to persuasiveness and strength ratings led me to merge these aspects into a single measure of how convincing the argument was, along with the influence rating.

Procedure and Measures

I retained one of the legal dilemmas (parole) used in Studies 1 and 2, and the measures and procedures remain similar. However, Study 3 introduces some notable changes. Participants receive one argument from either a human or an AI legal advisor. These arguments, different from the ones used in previous studies, have been pretested to ensure distinction between different quality levels.

Once the participants complete the consent procedure, they receive a brief introduction, tailored to their assigned Advisor condition:

> “*[The use of Artificial Intelligence (AI) as advisors in legal proceedings alongside human advisors is an intriguing possibility.]*
>
> *We present you with a challenging decision scenario (shortened from a real legal case) in which an [AI/legal] advisor proposes a certain decision. After the case, we will ask you about your impressions.”*

Participants proceed by reading the parole case and indicating their initial support for the decision using a slider bar, mirroring the procedure in Studies 1 and 2. Then, they are presented with an argument from the randomly assigned advisor (human or AI), who advocates for one side of the case (randomly assigned to be pro vs. con) with a certain argument quality (good or bad) and type (case-specific, statistical, or referencing another case), also randomly assigned. Following this, participants reassess their support for the decision and answer how convincing they find the argument and how much it influenced their support for the case (order is counterbalanced). Unlike previous studies, participants do not provide approval ratings. Next, they complete the same trust measure as in Study 2 for the type of advisor they are assigned to. Finally, participants provide demographic information and indicate their level of experience with AI and programming.

Design and Analysis

Overall, I have a fully between-subjects design with 2 (Advisor: AI vs. Human) x 2 (Argument Direction: Pro vs Con) x 2 (Argument Strength: Good vs Bad) x 3 (Argument Type: Case-specific, Statistical, or Referencing another case) conditions. I will employ ANOVAs to gather evidence for my hypotheses. In these analyses, updated support will assess the "cognitive" route, while trust measures, particularly moral trust, will evaluate the "affective" route. Our hypotheses will examine both main effects within the "case-at-hand evidence" condition as well as possible interactions.

Chapter 6: General Discussion

AI technologies are taking different forms and roles as they integrate into our everyday lives. Among these roles, their function as advisors holds particular promise and is gaining prominence across diverse domains, including healthcare, finance, and the military. Despite this increasing prevalence, the legal field has remained relatively unexplored in terms of understanding the dynamics of interactions between

humans and potential AI advisors. In this thesis, I examined how individuals respond to advice from human and AI advisors in legal cases with significant decision conflicts.

I first investigated the persuasive effects of legal advice in parol and insanity plea scenarios and found no immediate rejection of AI advisors, as the algorithm aversion literature would suggest for a morally significant domain. People were equally persuaded by the counsel provided by AI advisors as they were by human advisors. This effect persisted even when human and AI advisors were positioned in direct opposition to each other in Study 2, presenting conflicting advice on the legal dilemma at hand. However, people's approval and trust ratings showed a different trend: people consistently expressed lower levels of approval and trust in artificial advisors compared to their human counterparts despite being equally influenced by their advice. This divergence in approval and trust ratings, even in the face of equal persuasion effect, suggests a potential dissociation between central information processing, which involves content-based persuasion effects, and peripheral processing, which encompasses affective responses related to approval and trust.

In Chapter 1, I introduced the concept of separate processing paths within the persuasion literature. In this project, I apply this dual path framework to address the conflicting findings observed in the phenomenon of algorithm aversion and algorithm appreciation. Depending on various factors such as the nature of the task being performed (subjective or objective), the portrayal of the agent providing the advice (highlighting expertise or not), and the specific outcome being measured (agreement or trust), individuals may adopt either a central or peripheral processing route. Through the central route, individuals tend to engage in a more deliberative and analytical evaluation, which can lead to responses that are fair and perhaps even favorable towards AI. Conversely, the peripheral route often reveals a persistent unease or discomfort with AI, particularly in domains traditionally associated with human expertise, such as moral judgments (Bigman & Gray, 2018; Morewedge, 2022).

While the findings from this study do not necessarily threaten the potential use of AI advisors within the legal domain, it is imperative that these advisors present their arguments in a clear and comprehensible manner to ensure effective communication. Moreover, it is crucial to avoid triggering any discomfort or unease among individuals in their interactions with AI advisors. Even seemingly minor factors, such as the use of a machine voice or appearance (Dai et al., 2023; Pak et al., 2012), may inadvertently undermine central processing, leading to potential biases or hesitancy in taking advice from artificial advisors.

Future research should explore whether the publicly available evidence regarding the capabilities and qualities of AI, particularly within specific domains, could contribute to fostering greater trust and acceptance among users. Alternatively, it remains to be seen whether familiarity over time will naturally lead to increased liking and reliance on AI advisors. In the meantime, it is noteworthy that even esteemed institutions such as the U.S. Supreme Court are closely monitoring developments in this technological frontier, emphasizing the necessity for "caution and humility" in the use of AI (Roberts, 2023, p.5).

Limitations

While these studies provide new insights into human responses to AI, it is important to acknowledge their limitations. Even though I selected two distinct legal cases from a pool of six, generalizability to other legal scenarios remains a question. By design, I presented high-conflict scenarios with well-formed arguments (Studies 1 and 2), which made the legal advice compelling, regardless of which advisor conveyed it (Study 3 will investigate this limitation). I also did not provide information on the design of the artificial agent nor set an expectation of expertise for either advisor. Therefore, it is possible that AI might have been perceived as less persuasive if its expertise was in doubt, particularly if participants' expectations for AI-generated arguments were high (Renier et al., 2021).

Furthermore, participants in these studies were positioned as courtroom observers, tasked only with providing judgments about the legal decisions presented to them. They were not required to act on these judgments, nor were they subjected to any potential risks, such as financial loss. To address these limitations, future research should explore more realistic settings that require direct involvement, such as placing participants in the role of jurors, where they must make consequential decisions. Additionally, investigating scenarios involving costly decisions could provide further insights into how individuals respond to legal advice from both human and AI advisors in real-world contexts.

Additionally, I did not manipulate the anthropomorphism of the artificial advisors in these studies. In real-life settings, the manner in which AI advice is delivered could impact the outcome. If the non-human characteristics of the AI are highlighted during the interaction, individuals may become more hesitant to trust and use the advice provided. For example, if the AI's responses lack warmth or human-like qualities in their delivery, it could potentially diminish the perceived credibility and trustworthiness of the advice. Therefore, future research should consider the impact of anthropomorphism on the effectiveness of AI advisors in influencing decision-making processes to ensure the smoothest integration of the technology.

My focus was on AI advice, rather than AI decision-making, as the advisor role of AI is closer to reality. However, it is worth noting that people might pose higher standards (e.g., more transparency, higher accuracy, proven expertise, etc.) to endorse AI decision-makers than AI advisors. This is especially true in domains where AI has been found to violate fairness norms and engage in discriminatory practices (Burk, 2021; Dressel & Farid, 2018). Additionally, I specifically examined the effects of valid arguments conveyed by AI (even in low-quality arguments, there is no indication that the evidence is wrong), which seems to be an optimistic aim for now. While Study 3 results may shed light on how people discern well-constructed AI advice from badly constructed ones, further research should focus on understanding people's reactions to good *and* bad advice coming from artificial agents.

For the successful integration of AI technologies, people need to develop the ability to evaluate AI advice critically, just as they do between good and bad advice when it comes from humans. I hope that my thesis provided new knowledge on AI advisors and inspiration for future research.

References

Agrawal, A., Gans, J. S., & Goldfarb, A. (2019). Exploring the impact of artificial Intelligence: Prediction versus judgment. *Information Economics and Policy*, *47*, 1–6. https://doi.org/10.1016/j.infoecopol.2019.05.001

Al-Alawi, A. I., & Al-Mansouri, A. M. (2023). Artificial intelligence in the judiciary system of Saudi Arabia: A literature review. *2023 International Conference On Cyber Management And Engineering (CyMaEn)*, 83–87. https://doi.org/10.1109/CyMaEn57228.2023.10050929

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Avendaño, D., Sofia, C., Zapata, P., Portaluri, A., Maria Orlando, A. A., Avalos, P., Blandino, A., Ascenti, G., Cardona-Huerta, S., & Marino, M. A. (2023). Artificial Intelligence in Breast Imaging: A Special Focus on Advances in Digital Mammography & Digital Breast Tomosynthesis. *Current Medical Imaging*, *19*(8), 799–806. https://doi.org/10.2174/1573405619666221128102209

Bailey, P. E., Leon, T., Ebner, N. C., Moustafa, A. A., & Weidemann, G. (2023). A meta-analysis of the weight of advice in decision-making. *Current Psychology*, *42*(28), 24516–24541. https://doi.org/10.1007/s12144-022-03573-2

Belazoui, A., Telli, A., & Arar, C. (2022). An Adaptive Medical Advisor to Improve Diabetes Quality of Life. In S. Sedkaoui, M. Khelfaoui, R. Benaichouba, & K. Mohammed Belkebir (Eds.),

International Conference on Managing Business Through Web Analytics (pp. 259–268). Springer International Publishing. https://doi.org/10.1007/978-3-031-06971-0_19

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2021). Watch Me Improve—Algorithm Aversion and Demonstrating the Ability to Learn. *Business & Information Systems Engineering*, *63*(1), 55–68. https://doi.org/10.1007/s12599-020-00678-5

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151. https://doi.org/10.1016/j.obhdp.2006.07.001

Bower, A. H., & Steyvers, M. (2021). Perceptions of AI engaging in human expression. *Scientific Reports*, *11*(1), 21181. https://doi.org/10.1038/s41598-021-00426-z

Burk, D. L. (2021). Algorithmic legal metrics. *Notre Dame Law Review*, *96*(5), 1147–1204.

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

Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010). Detecting Management Fraud in Public Companies. *Management Science*, *56*(7), 1146–1160.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, *39*(5), 752–766. https://doi.org/10.1037/0022-3514.39.5.752

Chaiken, S., Liberman, A., & Eagly, A. H. (1989). Heuristic and systematic information processing within and beyond the persuasion context. In J. S. Uleman & J. A. Bargh (Eds.), *Unintended Thought* (pp. 212–252). Guilford.

Chugunova, M., & Sele, D. (2022). We and It: An interdisciplinary review of the experimental evidence on how humans interact with machines. *Journal of Behavioral and Experimental Economics*, *99*, 101897. https://doi.org/10.1016/j.socec.2022.101897

Clark v. Arizona (United States Supreme Court 2006). https://www.apa.org/about/offices/ogc/amicus/clark

Dai, Y., Lee, J., & Kim, J. W. (2023). Ai vs. human voices: How delivery source and narrative format influence the effectiveness of persuasion messages. *International Journal of Human–Computer Interaction*, *0*(0), 1–15. https://doi.org/10.1080/10447318.2023.2288734

Day, M.-Y., Cheng, T.-K., & Li, J.-G. (2018). AI Robo-Advisor with Big Data Analytics for Financial Services. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 1027–1031. https://doi.org/10.1109/ASONAM.2018.8508854

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2014). *Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err* (SSRN Scholarly Paper 2466040). https://doi.org/10.2139/ssrn.2466040

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580

Feng, B., & MacGeorge, E. L. (2010). The influences of message and source factors on advice outcomes. *Communication Research*, *37*(4), 553–575. https://doi.org/10.1177/0093650210368258

Goel, M., Tomar, P. K., Vinjamuri, L. P., Swamy Reddy, G., Al-Taee, M., & Alazzam, M. B. (2023). Using AI for Predictive Analytics in Financial Management. *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 963–967. https://doi.org/10.1109/ICACITE57410.2023.10182711

Goodwin, P., Sinan Gönül, M., & Önkal, D. (2013). Antecedents and effects of trust in forecasting advice. *International Journal of Forecasting*, *29*(2), 354–366. https://doi.org/10.1016/j.ijforecast.2012.08.001

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic)

and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law*, *2*(2), 293–323. https://doi.org/10.1037/1076-8971.2.2.293

Gsenger, R., & Strle, T. (2021). Trust, Automation Bias and Aversion: Algorithmic Decision-Making in the Context of Credit Scoring. *Interdisciplinary Description of Complex Systems*, *19*(4), 542–560. https://doi.org/10.7906/indecs.19.4.7

Hanson, A., Starr, N. D., Emnett, C., Wen, R., Malle, B. F., & Williams, T. (2024). *The Power of Advice: Differential Blame for Human and Robot Advisors and Deciders in a Moral Advising Context*.

Hou, Y., & Jung, M. (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. https://doi.org/10.1145/3479864

Hwang, T.-H., Lee, J., Hyun, S.-M., & Lee, K. (2020). Implementation of interactive healthcare advisor model using chatbot and visualization. *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, 452–455. https://doi.org/10.1109/ICTC49870.2020.9289621

Jones, O. T., Matin, R. N., van der Schaar, M., Prathivadi Bhayankaram, K., Ranmuthu, C. K. I., Islam, M. S., Behiyat, D., Boscott, R., Calanzani, N., Emery, J., Williams, H. C., & Walter, F. M. (2022). Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: A systematic review. *The Lancet. Digital Health*, *4*(6), e466–e476. https://doi.org/10.1016/S2589-7500(22)00023-1

Jones-Jang, S. M., & Park, Y. J. (2023). How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. *Journal of Computer-Mediated Communication*, *28*(1), zmac029. https://doi.org/10.1093/jcmc/zmac029

Kayande, U., De Bruyn, A., Lilien, G. L., Rangaswamy, A., & van Bruggen, G. H. (2009). How Incorporating Feedback Mechanisms in a DSS Affects DSS Evaluations. *Information Systems Research*, *20*(4), 527–546. https://doi.org/10.1287/isre.1080.0198

Kim, J., Merrill Jr., K., Xu, K., & Collins, C. (2023). My Health Advisor is a Robot: Understanding

Intentions to Adopt a Robotic Health Advisor. *International Journal of Human–Computer Interaction*, *0*(0), 1–10. https://doi.org/10.1080/10447318.2023.2239559

Kühne, K., Fischer, M. H., & Zhou, Y. (2020). The Human Takes It All: Humanlike Synthesized Voices Are Perceived as Less Eerie and More Likable. Evidence From a Subjective Ratings Study. *Frontiers in Neurorobotics*, *14*. https://www.frontiersin.org/articles/10.3389/fnbot.2020.593732

Law, T., & Scheutz, M. (2021). Trust: Recent concepts and evaluations in human-robot interaction. In C. S. Nam & J. B. Lyons (Eds.), *Trust in Human-Robot Interaction* (pp. 27–57). Academic Press. https://doi.org/10.1016/B978-0-12-819472-0.00002-2

Liu, Y., & Moore, A. (2022). A bayesian multilevel analysis of belief alignment effect predicting human moral intuitions of artificial intelligence judgements. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44). https://escholarship.org/uc/item/3v79704h

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research*, *46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Longoni, C., & Cian, L. (2022). Artificial intelligence in utilitarian vs. hedonic contexts: The "word-of-machine" effect. *Journal of Marketing*, *86*(1), 91–108. https://doi.org/10.1177/0022242920957347

Malle, B. F., Magar, S. T., & Scheutz, M. (2019). AI in the Sky: How People Morally Evaluate Human and Machine Decisions in a Lethal Strike Dilemma. In M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi, & E. E. Kadar (Eds.), *Robotics and Well-Being* (pp. 111–133). Springer International Publishing. https://doi.org/10.1007/978-3-030-12524-0_11

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice One For the Good of Many?: People Apply Different Moral Norms to Human and Robot Agents. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 117–124.

https://doi.org/10.1145/2696454.2696458

Malle, B. F., & Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. In C. S. Nam & J. B. Lyons (Eds.), *Trust in Human-Robot Interaction* (pp. 3–25). Academic Press. https://doi.org/10.1016/B978-0-12-819472-0.00001-0

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.2307/258792

Morewedge, C. K. (2022). Preference for human, not algorithm aversion. *Trends in Cognitive Sciences*, *26*(10), 824–826. https://doi.org/10.1016/j.tics.2022.07.007

Önkal, D., Goodwin, P., Thomson, M., Gönül, S., & 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. https://doi.org/10.1002/bdm.637

Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics*, *55*(9), 1059–1072. https://doi.org/10.1080/00140139.2012.691554

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

People v. Barnes (Appellate Division of the Supreme Court of New York, Fourth Department 1990). https://casetext.com/case/people-v-barnes-364

People v. Watson (Supreme Court of California 1981). https://law.justia.com/cases/california/supreme-court/3d/30/290.html

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 19, pp. 123–205). Academic Press. https://doi.org/10.1016/S0065-2601(08)60214-2

Prahl, A., & Swol, L. V. (2021). Out with the Humans, in with the Machines?: Investigating the Behavioral and Psychological Effects of Replacing Human Advisors with a Machine. *Human-Machine Communication*, *2*(1). https://doi.org/10.30658/hmc.2.11

Promberger, M., & Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making*, *19*(5), 455–468. https://doi.org/10.1002/bdm.542

Qiu, L., & Benbasat, I. (2009). Evaluating Anthropomorphic Product Recommendation Agents: A Social Relationship Perspective to Designing Information Systems. *Journal of Management Information Systems*, *25*(4), 145–182. https://doi.org/10.2753/MIS0742-1222250405

Renier, L. A., Schmid Mast, M., & Bekbergenova, A. (2021). To err is human, not algorithmic – Robust reactions to erring algorithms. *Computers in Human Behavior*, *124*, 106879. https://doi.org/10.1016/j.chb.2021.106879

Roberts, J. G. (2023, December 31). *2023 year-end report on the federal judiciary*. Chief Justice's Year-End Reports on the Federal Judiciary - Supreme Court of the United States. https://www.supremecourt.gov/publicinfo/year-end/2023year-endreport.pdf

Schecter, A., Bogert, E., & Lauharatanahirun, N. (2023). Algorithmic appreciation or aversion? The moderating effects of uncertainty on algorithmic decision making. *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–8. https://doi.org/10.1145/3544549.3585908

Schemmer, M., Hemmer, P., Kühl, N., Benz, C., & Satzger, G. (2022). *Should I Follow AI-based Advice? Measuring Appropriate Reliance in Human-AI Decision-Making* (arXiv:2204.06916). arXiv. https://doi.org/10.48550/arXiv.2204.06916

Schilbach, L., Eickhoff, S. B., Schultze, T., Mojzisch, A., & Vogeley, K. (2013). To you I am listening: Perceived competence of advisors influences judgment and decision-making via recruitment of the amygdala. *Social Neuroscience*, *8*(3), 189–202. https://doi.org/10.1080/17470919.2013.775967

Sniezek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision making. *Organizational Behavior and Human Decision Processes*, *62*(2), 159–174. https://doi.org/10.1006/obhd.1995.1040

Sniezek, J. A., & Van Swol, L. M. (2001). Trust, Confidence, and Expertise in a Judge-Advisor System.

Organizational Behavior and Human Decision Processes, *84*(2), 288–307. https://doi.org/10.1006/obhd.2000.2926

Sober, E. (1992). Screening-Off and the Units of Selection. *Philosophy of Science*, *59*(1), 142–152.

Soni, K., & Hasija, Y. (2022). *Artificial Intelligence Assisted Drug Research and Development | IEEE Conference Publication | IEEE Xplore*. https://ieeexplore.ieee.org/document/9753179

Stein, J.-P., Appel, M., Jost, A., & Ohler, P. (2020). Matter over mind? How the acceptance of digital entities depends on their appearance, mental prowess, and the interaction between both. *International Journal of Human-Computer Studies*, *142*, 102463. https://doi.org/10.1016/j.ijhcs.2020.102463

Straßmann, C., Eimler, S., Arntz, A., Grewe, A., Kowalczyk, C., & Sommer, S. (2020). *Receiving Robot's Advice: Does It Matter When and for What?* https://link.springer.com/chapter/10.1007/978-3-030-62056-1_23

Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters*, *10*(1–2), 1–10. https://doi.org/10.1049/htl2.12039

Thurman, N., Moeller, J., Helberger, N., & Trilling, D. (2019). My Friends, Editors, Algorithms, and I. *Digital Journalism*, *7*(4), 447–469. https://doi.org/10.1080/21670811.2018.1493936

Ullman, D., & Malle, B. F. (2018). What does it mean to trust a robot? Steps toward a multidimensional measure of trust. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 263–264). ACM. http://doi.acm.org/10.1145/3173386.3176991

Ullman, D., & Malle, B. F. (2023). *MDMT: Multi-Dimensional Measure of Trust (v2)*. Brown University. https://research.clps.brown.edu/SocCogSci/Measures/MDMT_v2 (2023).pdf

Wang, N. (2020). "Black Box Justice": Robot Judges and AI-based Judgment Processes in China's Court System. *2020 IEEE International Symposium on Technology and Society (ISTAS)*, 58–65. https://doi.org/10.1109/ISTAS50296.2020.9462216

Xu, J. Y., Branch, C., & Wang, Y. (2020). A Methodology and Experiments towards Autonomous

Decision Making. *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 1027–1032. https://doi.org/10.1109/SMC42975.2020.9283351

You, S., Yang, C. L., & Li, X. (2022). Algorithmic versus Human Advice: Does Presenting Prediction Performance Matter for Algorithm Appreciation? *Journal of Management Information Systems*, *39*(2), 336–365. https://doi.org/10.1080/07421222.2022.2063553

Zeng, Q., Klein, C., Caruso, S., Maille, P., Laleh, N. G., Sommacale, D., Laurent, A., Amaddeo, G., Gentien, D., Rapinat, A., Regnault, H., Charpy, C., Nguyen, C. T., Tournigand, C., Brustia, R., Pawlotsky, J. M., Kather, J. N., Maiuri, M. C., Loménie, N., & Calderaro, J. (2022). Artificial intelligence predicts immune and inflammatory gene signatures directly from hepatocellular carcinoma histology. *Journal of Hepatology*, *77*(1), 116–127. https://doi.org/10.1016/j.jhep.2022.01.018

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305. https://doi.org/10.1145/3351095.3372852