Dissociated Responses to AI: Persuasive But Not Trustworthy?

Zeynep Aydin (zeynep_aydin@brown.edu)
Department of Cognitive, Linguistic, and Psychological Sciences, Brown University,
190 Thayer Street. Providence RI 02906 USA

Bertram F. Malle (bfmalle@brown.edu)
Department of Cognitive, Linguistic, and Psychological Sciences, Brown University,
190 Thayer Street. Providence RI 02906 USA

Abstract

Empirical work on people's perceptions of AI advisors has found evidence for both "algorithm aversion" and "algorithm appreciation." We investigated whether these differing reactions stem from two different paths of processing: assessing the content of the advice and evaluating the source (AI vs. human advisor). In two survey studies, people were as strongly persuaded by the advice of an AI as that of a human advisor; nonetheless, people's approval of and trust in the AI advisor was consistently lower. This pattern of dissociation suggests that algorithm aversion and algorithm appreciation can occur at the same time, but along different response paths.

Keywords: Artificial Intelligence; Psychology; Human-computer interaction; Reasoning; Social cognition

As society adapts to emerging artificial intelligent agents, questions about their role—whether as tools, assistants, or partners—remain open to debate. In this debate, people's psychological responses to artificial agents play a critical role. Understanding how individuals perceive and interact with AI is essential not only for designing socially acceptable agents but also for gaining insights into fundamental psychological processes of social cognition, moral psychology, and social influence. Some work on future artificial agents has studied agents as decision-makers (Gsenger & Strle, 2021; Malle et al., 2015; Sen et al., 2023; Xu et al., 2020), but researchers are increasingly examining artificial *advisors* (Hanson et al., 2024; Straßmann et al., 2020), which is perhaps the more imminent role of AI (Belazoui et al., 2022; Goel et al., 2023; Hwang et al., 2020; Kim et al., 2023). Therefore, our focus in this paper will be on people's perceptions of AI advisors.

Empirical work on these perceptions has yielded contradictory findings often labeled "algorithm aversion" and "algorithm appreciation." Algorithm aversion refers to people's reluctance to trust the advice of AI compared to human advice (Dietvorst et al., 2014; Jones-Jang & Park, 2023; Promberger & Baron, 2006). Conversely, algorithm appreciation refers to the greater reliance on AI over human advice (Logg et al., 2019; Schecter et al., 2023; Thurman et al., 2019). Researchers have begun to identify the conditions that give rise to these contrasting stances, such as task type (Castelo et al., 2019; Longoni & Cian, 2022) or portrayal of expertise (Hou & Jung, 2021). Specifically, Hou and Jung (2021) proposed a comprehensive explanation that accounts for a considerable number of findings: In studies that describe an artificial agent as powerful and competent, people appreciate and accept its advice; in studies that describe it as less competent than humans, people reject the advice.

This notion of expert power to explain the contradictory findings may still be insufficient, however. Recent studies suggest that people's processing of AI-generated information can dissociate from their affect, liking, or trust toward the AI (Bower & Steyvers, 2021; Liu & Moore, 2022; Renier et al., 2021). The suggestion of dissociated responses implies a distinction between content-based processing, which involves a thoughtful assessment of the offered advice, and source evaluation, which is a more automatic response of approving of, liking, or trusting the source (Chaiken et al., 1989; Petty & Cacioppo, 1986). Our studies use this distinction to explore whether individuals have dissociated responses to AI versus human advisors by processing the content of the advice through one pathway, potentially changing their opinions, while simultaneously evaluating the source of the advice through a separate pathway.

Studies on AI advisors have primarily explored settings related to event forecasting (e.g., Logg et al., 2019; Önkal et al., 2009) and medical diagnosis (e.g., Longoni et al., 2019; Promberger & Baron, 2006). Only a few studies have examined legal proceedings, despite growing interest in real-world applications of AI in the law (Al-Alawi & Al-Mansouri, 2023; Angwin et al., 2016; Roberts, 2023; Wang, 2020). Therefore, we investigated people's responses to legal advisors and assessed the separate paths of content-based processing (persuasive message effects) and source evaluation (AI vs. human advisor effects).

In two studies, participants encountered legal dilemmas where either one of two decisions (e.g., granting parole or not) could be reasonably supported. After reading about the dilemma and indicating their initial stance (baseline support), people received an argument from either a human or AI legal advisor, favoring one or the other decision. Then participants indicated their updated 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). Additionally, we captured source evaluation responses in people's rated approval of the AI/ human advisor (Studies 1 and 2) and perceived trustworthiness (Study 2). Thus, we were able to examine potential "aversion," "appreciation," or

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

even-handed responses to AI along two routes: (1) content-based persuasion effects and/or (2) source-based approval and trust effects elicited by human and AI advisors.

We pretested numerous cases to design credible legal dilemmas and selected two with equal support for each of the two possible decisions. The type of advisor was manipulated merely by changing the expression "legal advisor" to "AI legal advisor." Each advisor advocated for one or the other decision in the case. In Study 1, participants read about both legal cases featuring the same type of advisor but received a pro-argument in one case and a con-argument in the other. In Study 2, participants read about only one of the two cases but with opposing arguments from a human and an AI advisor.

Study 1

Methods

Participants We aimed for 100 participants in each of the AI/human advisor conditions to detect an effect size of $d \geq 0.40$. Using the online crowdsourcing website *Prolific*, we recruited 216 participants. 20 cases failed a bot check, leaving 196 for analysis (mean age = 39.7; 90 identified as female, 101 as male, and 5 as nonbinary or genderqueer). Participants received $1.00 for completing the 6-minute survey.

Stimulus Selection We first designed six candidate narratives derived from real legal cases (*Clark v. Arizona*, 2006; *People v. Barnes*, 1990; *People v. Watson*, 1981), each involving a difficult decision (e.g., granting parole or not, lenient vs. harsh sentence) where both options could be reasonably defended. In a pretest sample, 162 respondents indicated on a 100-point scale how much they supported one or the other decision (with opposing decisions labeling the poles). We selected two cases for which mean support ratings were closest to the 50 mark. The chosen scenarios were a *parole* case ($M = 55.4$) and an *insanity plea* case ($M = 49.9$).

Procedure and Measures After the consent procedure, participants received a brief introduction about the use of [AI] legal advisors in legal proceedings (see Supplementary Materials at https://bit.ly/LegAIAdv). Then they read one *Case* (parole or insanity plea, counterbalanced) and indicated baseline support for the decision. 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." Next, they received a recommendation from their assigned *Advisor* (AI vs. human) with a randomly assigned *Argument direction* (e.g., for vs. against parole). After that, participants rated their updated support as well as their overall approval of the advisor ("How much do you approve of the AI advisor giving this particular advice?"), also on a 0-100 scale. Then, 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 the same measures of baseline and updated support as well as advisor approval for this case. Finally, they provided information about age, gender, AI knowledge, and programming experience.

Results

At baseline, people leaned slightly toward granting parole ($M = 54.6$, $SD = 27.8$), diverging from 50, $t(195) = 2.30$, $p = .023$; and they leaned slightly against accepting the insanity plea ($M = 44.4$, $SD = 29.8$), $t(195) = -2.60$, $p = .01$. Support ratings in the two cases were largely independent, $r = .167$ ($p = .019$). To examine robustness, we analyzed each case separately.

Persuasion Effects. We first analyzed the "persuasion effect"—how much participants changed their support in the direction of the presented (pro- or con-)arguments (see Figure 1). We used a mixed-design ANOVA with between-subjects factors Advisor (AI, human) and Argument (pro vs. con) and within-subject factor Support change (baseline to updated judgment). The persuasion effect corresponds to the Support change × Argument interaction; a moderation of this persuasion effect by AI vs. human is captured by the three-way interaction of Support change × Argument × Advisor.

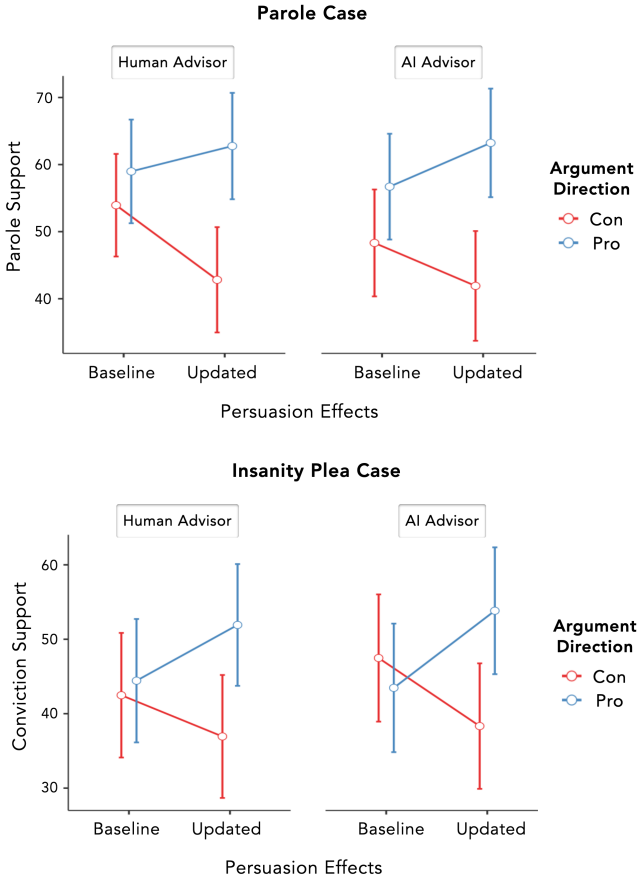

Figure 1: Study 1 shows highly similar persuasion effects (support change in response to arguments) for AI and Human advisor in the parole case (top) and the insanity plea case (bottom). Error bars are 95% CIs.

For the parole case, the persuasion effect was substantial, $F(1, 192) = 65.4$, $p < .001$, $\eta^2 = 25.4\%$. Specifically, the pro-argument increased support (+5.1 points on a 0-100 scale), and the con-argument decreased support (-8.9 points). No three-way interaction emerged with type of Advisor, $F < 1$, $\eta^2 < 0.1\%$. A follow-up Bayesian ANOVA revealed strong evidence against such an interaction, $BF_{10} = 0.25$.

For the insanity plea case, the persuasion effect was also strong, with pro-arguments increasing (+ 8.8) and con-arguments decreasing support (-7.3), $F(1, 192) = 47.9$, $p < .001$, $\eta^2 = 20\%$. No significant interaction emerged with Advisor, $F(1, 192) = 1.9$, $p = .17$, $\eta^2 = 1\%$. A Bayesian ANOVA revealed inconclusive evidence against such an interaction, $BF_{10} = 0.51$.

Approval. In contrast to equally strong persuasion effects for AI and human advisor, people approved less of the AI than the human (see Figure 2). In the parole case, the AI received approval of $M = 54.1$, while the human advisor received $M = 62.0$, $F(1, 192) = 3.48$, $p = .064$, $\eta^2 = 2\%$. In the insanity plea case, too, the AI advisor received lower approval ratings ($M = 49.1$) than the human ($M = 59.0$), $F(1, 192) = 5.47$, $p = .02$, $\eta^2 = 2.8\%$. These approval differences did not interact with Argument direction ($ps > .28$, $\eta^2s < 1\%$), so diverging approval was caused by the type of *advisor*, irrespective of the argument the advisor made.

Approval ratings were uncorrelated with support change scores, both for parole (full sample $r = -.12$, Human $r = -.17$, AI $r = -.11$) and for insanity plea (full sample $r = .05$, Human $r = .02$, AI $r = .12$), all $ps > .11$.

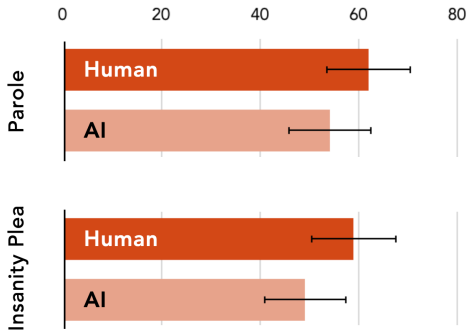

Figure 2: Study 1 shows greater approval for the human advisor than for the AI advisor in the parole case (left) and the insanity plea case (right). Error bars are 95% CIs.

Discussion

The two legal cases elicited the intended decision conflict, as people's support was roughly centered on the scale midpoint. But when exposed to an advisor's argument for one or the other side, people were persuaded and shifted their support in the direction of the argument. Importantly, this persuasion effect did not differ between human and AI advisors, indicating no sign of algorithm aversion or algorithm appreciation for content-based information processing. By contrast, participants approved less of the AI advisor than of the human advisor, irrespective of argument direction, indicating affective reservations toward the AI as a source. These reservations were uncorrelated with support change, which suggests that responses to AI (vs. human) advisors occur along two potentially independent routes of processing: content-based and source-based.

However, the equal-sized persuasion effect for human and AI may have arisen because people considered only one agent (and, in a given case, only one argument) at a time. To examine this possibility, Study 2 invited people to directly compare the two agents giving opposing arguments. This situation may allow affective reservations of AI to distort people's message content processing. For example, people might construe a given argument expressed by a human as more reasonable than the opposing argument expressed by the AI, regardless of which argument is offered by whom (Dai et al., 2023; Feng & MacGeorge, 2010).

"Advice utilization" (what we labeled persuasion effect) is often regarded as a metric for trust—an intuitive sense of trusting the advisor (Bonaccio & Dalal, 2006; Sniezek & Van Swol, 2001). This perspective, however, equates the construct of trust with reliance (i.e., utilization). Trust is a subjective expectation that the advisor is capable, reliable, sincere, has integrity, and so on (Malle & Ullman, 2021; Mayer et al., 1995), whereas reliance is the decision to use the advice (whether out of trust, convenience, or necessity). Recent studies have shown that trust is multi-dimensional, encompassing both performance trust (expecting the agent to be competent and reliable) and moral trust (expecting the agent to be ethical, sincere, and benevolent); and this multi-dimensionality applies to trusting both humans and machines (Malle & Ullman, 2021). Measuring trust in this multi-dimensional way allows us to capture whether aversion (or appreciation) of AI is grounded in judgments of performance or in judgments of moral capabilities. Some authors have claimed that performance considerations are a driving force of AI aversion (Hou & Jung, 2021; Parasuraman & Riley, 1997), whereas others have argued that people specifically reject AI in moral domains (Bigman & Gray, 2018), and many legal matters have moral undertones. Therefore, in Study 2, we added a multi-dimensional trust measure to assess whether trust (performance or moral) would operate similarly to approval ratings, namely showing lower levels for an AI than a human advisor, irrespective of their arguments. This difference in evaluating the source should contrast with the content-based processing in which AI and human advisors have equal persuasive effects.

Study 2

Participants received counsel from both agents, a human, and an AI advisor, giving opposing arguments. We evaluated the extent to which participants were persuaded by each of the advisors. We counterbalanced other factors, including the pairing of each agent with a specific argument and the order in which agents presented their arguments.

Methods

Participants We aimed for 100 participants in each of the AI/human advisor conditions to detect an effect size of d ≥ 0.40. Again, using Prolific, we recruited 247 participants. 11 participants did not enter any responses, and 23 responses failed the bot check, leaving 211 valid responses (mean age = 41.6; 126 identified as female, 83 as male, and 2 as nonbinary or genderqueer). Each participant received $1.40 in compensation for completing the 7-minute survey.

Procedure and Measures We used the same two legal dilemmas as in Study 1, and the measures and procedures were similar as well. However, in Study 2, participants received only one case (parole or insanity plea) but two arguments, one from a human legal advisor and the opposite one from an AI legal advisor. The arguments were the same ones as in Study 1, pretested to have comparable persuasive strength, resulting in a strong advice opposition between human and AI advisors.

After indicating their baseline support judgments about their assigned case, participants received the argument from the first advisor (human or AI, counterbalanced), who favored one side in the case (randomly assigned to be pro vs. con). Participants made their first updated support judgment in response to this first argument and expressed their approval of the first advisor. Then the other advisor presented their argument, favoring the opposite side, and participants gave their second updated support judgment in response to this argument and their approval of the second advisor. Next, participants completed a trust measure for one of the advisors and then a parallel form of that trust measure for the other advisor (both advisor and forms were counterbalanced). Finally, participants provided demographic information and indicated their level of experience with AI and programming.

We measured trust with the Multi-Dimensional Measure of Trust (MDMT v2, Ullman & Malle, 2018, 2023). The MDMT's underlying theoretical model conceptualizes trust as a multi-dimensional set of expectations about the agent's trustworthy dispositions (Malle & Ullman, 2021). It is designed to assess people's trust both in human and artificial agents and does so along two major factors: Performance trust and Moral trust. Because participants had to complete the trust measure twice, for each of the advisors, we used the parallel 10-item short forms of the MDMT, which have entirely distinct items but are highly correlated.

Design and Analysis Each participant saw only one of two *Cases* (parole vs. insanity plea) but made judgments about both the human and the AI advisor (in counterbalanced order). However, because people always saw opposing arguments (e.g., pro parole by the human advisor and con parole by the AI), the *Argument* and *Advisor* factors were not crossed. We therefore adopted a linear mixed-effects approach that elegantly handles such dependencies and analyzed the two legal cases separately to examine robustness.

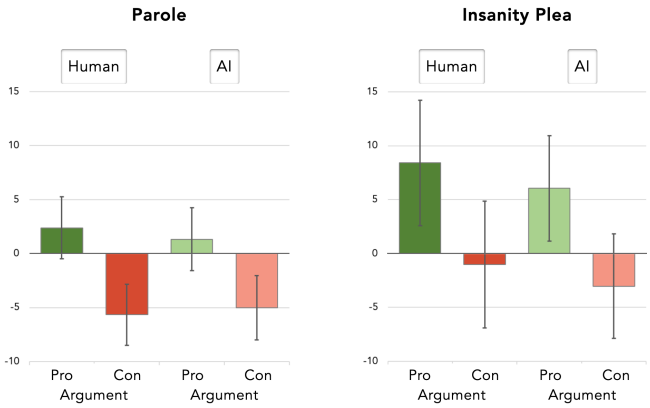

Figure 3: Study 2 shows similar persuasion effects (support change from baseline) for AI and human advisors in the parole case (left) and the insanity plea case (right). Error bars are 95% Cis from raw data.

Results

At baseline, Participants leaned against granting parole (M = 44.7, SD = 26.6), diverging from 50, $t(104) = -2.04$, $p = .043$, but were nearly neutral toward accepting the insanity plea (M = 47.0, SD = 28.8), $t(105) = -1.09$, $p = .28$.

Persuasion effects As in Study 1, we measured the "persuasion effect" as participants' change in support from baseline to updated judgment. Here, we had two updated judgments, one for each advisor, and in response to either pro- or con-arguments (but always in opposing directions between advisors). We thus predicted support change in a linear mixed effects model from Advisor (AI, human), Argument direction (pro or con), and Advisor order (human or AI first). We analyzed each case separately (see Figure 3).

In parole, the expected persuasion effects emerged, showing that people who received a pro-argument changed upwards by 1.83 points while those who received a con-argument changed downward by 4.84 points, $F(1, 101) = 31.3$, $p < .001$, $d = 0.58$. This persuasion effect did not significantly interact with Advisor, $F(1, 101) < 1$, $p = .47$ (see Fig. 3, left). A follow-up Bayesian analysis confirmed strong evidence against such an interaction, $BF_{10} = 0.16$.

In the insanity plea case, people who received a pro-argument changed upward by 7.29 points and those who received a con-argument changed downward by 2.52 points, $F(1, 102) = 20.6$, $p < .001$, $d = 0.49$. This persuasion effect again did not interact with Advisor, $F(1, 101) < 1$, $p = .93$ (see Fig. 3, right), confirmed by a follow-up Bayesian analysis, $BF_{10} = 0.17$.

Approval Approval judgments were distinct from the persuasion effects (see Figure 4). In the parole case, people approved of the human advisor more (M = 62.1) than of the AI advisor (M = 50.6), $F(1, 200) = 8.6$, $p = .019$, $d = 0.41$.

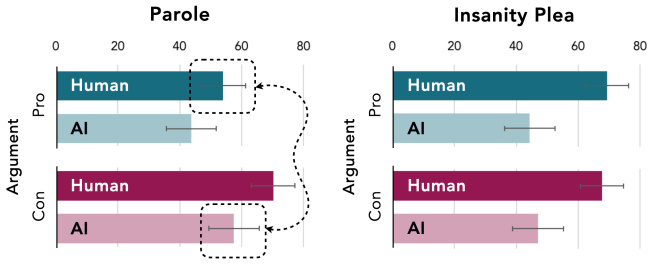

Figure 4: Study 2 shows consistently higher approval for the human than the AI advisor for the same arguments (pro or con), both for the parole case (left) and the insanity plea case (right). The dotted lines mark a seeming exception (see text for interpretation). Error bars are 95% CIs.

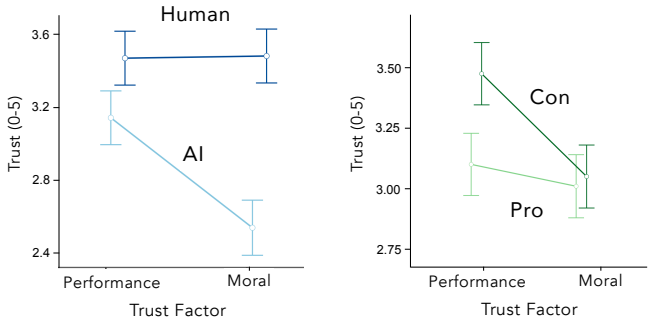

Figure 5: Left panel shows higher trust ratings in the human advisor than the AI advisor in Study 2, especially for moral trust. Right panel shows stronger effects on trust from con arguments than pro arguments for performance trust.

In the insanity plea case, too, people approved of the human advisor more (M = 68.7) than of the AI advisor (M = 45.4), $F(1, 102) = 24.0$, $p < .001$, $d = 0.82$. However, in the parole case, there was one condition in which human and AI advisor received nearly equal approval ratings (see dotted markings in Figure 4, left panel). This was because, in the parole case, the con-arguments received overall higher approval than the pro-arguments, $F(1, 200) = 14.8$, $p < .001$, so in the condition in which the AI's con-argument was pitted against the human's pro-argument, the AI approval deficit was balanced out by the pro-argument deficit. This seeming exception to the lower AI approval did not emerge in the insanity plea case because people had no preference for the pro- or con-arguments. Thus, the AI advisor received less approval than the human advisor, controlling for argument preference. In neither case was there any order effect of which advisor offered their argument first.

Trust Finally, we analyzed trust in the two advisors, which was measured after people had made all support and approval judgments. Because of the multiple within-subject dependencies, we chose a mixed-effects model with subject as random effect. There were no significant or noteworthy differences between the two cases, so we report the aggregated findings. The first result was that trust in the human advisor was considerably higher overall, $F(1, 618) = 96.7$, $p < .001$, and especially in Moral trust, $F(1, 618) = 22.7$, $p < .001$. As the left panel of Figure 5 shows, while people had only slightly higher Performance trust in the human (M = 3.47) than the AI advisor (M = 3.14), they had substantially higher Moral trust in the human (M = 3.48) than the AI advisor (M = 2.54). In addition, we found that con-arguments somewhat increased trust, $F(1, 618) = 6.5$, $p = .012$, but primarily for Performance trust, interaction $F(1, 408) = 19.1$, $p < .001$.

Additionally, we explored possible effects of order of presentation of the trust measure. Indeed, the AI's trust deficit was weaker when AI trust ratings came first, $F(1, 612) = 11.3$, $p < .001$, but for both Moral trust and Performance trust. Participants who expressed trust in the AI advisor first might have focused more on the positive aspects of AI, whereas those who considered AI second may have focused more on what it lacks compared to human advisors.

Relationship among measures Finally, we examined how the three main measures related to each other. We see in Table 1 that content-based support update is orthogonal to the more affective agent variables of approval and trust, and that is true 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 in any way. However, approval and trust are closely related. The AI's lower approval shrinks from an effect of η^2 = 11.1% to 5.7% when controlling for the AI Performance trust deficit, and it shrinks to 0% when controlling for the AI Moral trust deficit. Hence, the disapproval of AI is not entirely one of perceived competence limitations but primarily one of perceived shortcomings in the moral domain. The reverse analysis, controlling for approval in examining the AI trust deficits, shows that the overall trust deficit shrinks from 19% to 10%, but the greater Moral than Performance trust deficit is unchanged, as is the mitigation of performance trust when the AI offers a con-argument (Figure 5). Thus, approval judgments appear to be a reflection of trust concerns more than trust concerns are a reflection of approval judgments.

Table 1: Correlations among main measures in Study 2

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

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

Discussion

Study 2 directly pitted human and AI advisor arguments against each other in the same legal scenarios. We measured both changes in content-based support judgments resulting from the arguments as well as approval and trust toward the advisors. In line with Study 1, arguments in favor or against the legal decisions persuaded participants to shift their judgments in the direction of the arguments, regardless of whether they were expressed by a human or AI advisor. But, just as in Study 1, participants approved of the AI consistently less than of the human advisor. Trust, too, indicated an AI deficit, mildly at the level of Performance trust (competence, reliability) and substantially at the level of Moral trust (integrity, transparency, benevolence). Thus, the pattern of results suggests a dissociation between content-based persuasion (equally strong between human and AI) and agent-directed sentiments (greater approval and trust for human than AI).

Unfavorable sentiments toward the AI advisor were much weaker for Performance trust. We may speculate that Performance trust more closely measures the advisor's argument strength, whereas moral trust reflects more of the general affective reservations people have about AI. In addition, we saw that AI approval increased (in the parole case) when the AI offered advice for the decision that was slightly more popular and the human advisor offered the less popular advice. Here, the (affective) advisor effect was balanced out by the (cognitive) argument effect. Thus, despite the overall trend of dissociation between cognitive (content-based) processing and affective (source-based) processing, the two processing paths can interact.

General Discussion

AI technologies are increasingly used as advisors in such domains as healthcare, finance, and the military. Focusing on the less-explored legal domain, we examined how individuals respond to advice from human and AI advisors in legal cases with significant decision conflicts.

Our findings revealed that people were just as persuaded by AI as by human counsel. This persuasion effect remained equally strong even when human and AI gave opposite advice. However, people consistently expressed lower approval for and less trust in artificial agents, and this pattern was uncorrelated with the persuasion effects. These findings suggest that people's consideration of the advice contents and evaluation of the source may be dissociated.

The notion of separate paths of processing is reminiscent of a well-known distinction in the persuasion literature between central processing (deliberative consideration of message contents) and peripheral processing (relying on affect and heuristics, including source effects). This distinction provides a framework for reconciling seemingly contradictory findings of algorithm aversion and algorithm appreciation. Depending on a study's task (e.g., subjective or objective), agent description (salient expertise or not), and measured response (e.g., agreement or trust), people may use more of a central or peripheral processing path. The central path appears to allow for even-handed, sometimes even AI-favoring responses (e.g., for analytic tasks); the peripheral path reveals persistent discomfort with AI, especially in domains experienced to be uniquely human (e.g., moral judgments; Bigman & Gray, 2018; Morewedge, 2022).

The present results hold promise for the potential use of AI advisors in the legal domain, provided the advisors offer clearly communicated arguments and do not trigger people's discomfort with AI. However, even just a machine voice or look may undermine central processing (Andrews & Shimp, 1990; Dai et al., 2023; Pak et al., 2012). Likewise, people appear to have reasonable trust in the AI's capacity and reliability (Performance trust) but are much more reluctant to trust the AI as a morally competence agent. Further research should examine whether publicly available evidence for the qualities of AI in certain domains (e.g., extensive data analysis like in finance) might foster trust, at least additional Performance trust, or whether we must wait until familiarity and consistent performance breed liking. For now, even the U.S. Supreme Court pays close attention to this technological frontier but contends that "any use of AI requires caution and humility" (Roberts, 2023, p. 5).

Limitations

Our studies have clear limitations. We focused on the relatively understudied legal decision context with only two legal cases (preselected from a pool of six), so the generalizability to other legal cases and domains remains a concern. By design, we presented high-conflict scenarios with well-formed arguments, which made the legal advice compelling, regardless of which advisor conveyed it. If the arguments were weaker, AI might become less persuasive than humans, especially if people's expectations for AI arguments are high (Renier et al., 2021). Participants were placed in the role of courtroom observers who witnessed recommendations and reported only their judgments about what the legal decision should be; they did not have to act on those judgments and were not at any risk (e.g., loss of compensation). Future work needs to explore more realistic, costly decisions with more direct involvement (e.g., as jurors or loan recipients).

We focused on AI advice rather than AI decision-making, in part because the advisor role of AI is closer to reality. However, endorsing AI decisions is likely to have a higher threshold of acceptance. That is particularly true for domains where AI has been shown to violate norms of fairness and to engage in discrimination (Burk, 2021; Dressel & Farid, 2018). Our specific aim was to examine the effects of valid arguments conveyed by AI, which may, for now, be an overly optimistic aim. At the same time, for the successful integration of future AI advisors, it is important that people do not reject potentially valuable information merely because it comes from AI. People will need to discern good AI advice from bad AI advice, just as they need to do so when receiving advice from humans.

References

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

Andrews, J. C., & Shimp, T. A. (1990). Effects of involvement, argument strength, and source characteristics on central and peripheral processing of advertising. *Psychology & Marketing*, *7*(3), 195–214. https://doi.org/10.1002/mar.4220070305

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

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

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.

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.

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

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

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

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-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

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

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., 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
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
Sen, S., Kadam, S., & Ravi Kumar, V. V. (2023). Role of Artificial Intelligence-Enabled Recruitment Processes in Sourcing Talent. *2023 6th International Conference on Information Systems and Computer Networks (ISCON)*, 1–5. https://doi.org/10.1109/ISCON57294.2023.10112009
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
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
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