The influence of value similarity on psychological and behavioral trust in biased AI advice

Hendrikx, Yannic

Award date:
2023

Eindhoven, 31-1-2023

The influence of value similarity on psychological and behavioral trust in biased AI advice

by Yannic Hendrikx

identity number 1026076

in partial fulfilment of the requirements for the degree of

Master of Science
in Human Technology Interaction

Supervisors:
Dr. Gerrit Rooks
Prof. Dr. Chris Snijders

Abstract

Collaboration between AI and humans has the potential to improve decision-making accuracy and consistency. However, people are often reluctant to accept advice from AI. One of the leading causes of this low acceptance is a lack of trust. Past studies found that the alignment of moral values between a person and an AI positively affects trust in recommendations. It is proposed that people might also perceive inaccuracies in AI recommendations as value-laden judgments, especially in ethical or emotional decision-making situations. The current study examines how biased AI advice that aligns or misaligns with a user's preexisting values impacts trust and compliance. Additionally, differences in trust between AI and human advisors, trust development over time, and value attribution to AI are investigated.

To explore these questions an online experiment was conducted. Participants were tasked with determining prison sentences based on short descriptions of criminal offenses. They received advice from a fictional AI during the task. The AI's bias was expressed through its recommended prison sentences. These were either consistently lower or higher than those of a real-world judge. Self-reported trust was measured dynamically throughout the experiment. Adoption was measured based on how much participants adjusted their decisions to match the advice. The results show that an AI advisor that aligns with a user's values is trusted more. Furthermore, value similarity positively influenced the adoption of advice. Contrary to expectations, the results showed no difference in trust or adoption of advice between AI and human advisors. Similarly, no clear evidence of trust development over time was found. There were no significant differences between initial trust and later trust. This study concludes that user values and their possible influence on trust and compliance should be considered when developing advisory AI.

Contents

Introduction

As a result of advances in data collection and machine learning, the development and use of (assistive) decision-making algorithms have been increasing (EPRS, 2019). Decision-making algorithms are used in healthcare, finance, marketing, recruitment, and criminal justice. These systems can improve decision-making accuracy and efficiency when collaborating with humans (Grace et al., 2018). Algorithms can outperform human (expert) decision-makers in various domains (Grove et al., 2000). Furthermore, algorithms may provide solutions to current problems of human bias and prejudice in decision-making processes (Chiao, 2018).

Despite the potential benefits, people are reluctant to work with and accept recommendations from algorithms (Burton et al., 2020). One of the leading causes of the low acceptance of advisory algorithms is a lack of trust (Dzindolet, 2003; Lee, 2018). Prior research has identified numerous factors responsible for this lack of trust (Schaefer et al., 2016). For example, digital literacy (Logg, Minson & Moore, 2019), prior use by others (Alexander, Blinder & Zak, 2018), and task complexity (Fan et al., 2020). Particular attention has gone to the influence of errors on trust (Dietvorst, Simmons & Massey, 2015; Dietvorst et al., 2016). Inaccuracies in algorithmic advice lead to a loss of trust. Furthermore, trust lost due to inaccurate advice is hard to regain (Langer et al., 2022). People are less forgiving of erring algorithms than humans (Renier, Mast & Bekbergenova, 2021).

Proposed solutions to increase or regain trust are improved transparency and explainability (Shin, 2021), greater user decision autonomy (Dietvorst et al., 2016), and user expectation management (Goodyear et al., 2016). However, improvements in trust remain limited and are often constrained to specific situations (Burton et al., 2020). For instance, increased user autonomy results in more protracted decision-making, restricting this solution to situations that allow for ample collaboration between user and algorithm. Better calibration of expectations through increasing knowledge of an algorithm's functioning improves trust. However, requiring training and education before the use of a system is costly and time-consuming. Simplifying algorithmic processes to improve interpretability comes at the cost of system performance. Furthermore, detailed explanations of the system's reasoning are unfeasible for complex algorithms.

Prior research on AI (artificial intelligence) has shown that humans can perceive machine systems to have, and act based on moral values (Yokoi, Eguchi, Fujita & Nakayachi, 2021). People are more inclined to trust those they perceive as having similar values (Sitkin & Roth, 1993). Earle & Cvetkovich (1995) refer to this as salient value similarity (SVS). Value similarity refers to sharing one or more social values (Beilmann & Lilleoja, 2015). Value similarity positively influences trust in decisions made by humans (Siegrist, Cvetkovich & Roth, 2000). A positive influence of value-similarity on trust in decision-making systems was found by Mehrotra, Jonker & Tielman (2021). Participants interacted with five AI agents with distinct value profiles in their research. The AI agents provided recommendations in a hostage extraction scenario. Recommendations that aligned with the participants' values were trusted more. Yokoi & Nakayachi (2021) found greater trust in an AI-controlled vehicle that expressed similar moral values to the user. They used a modified version of the classic trolley problem to assess value similarity. Participants witnessed an autonomous vehicle make a utilitarian or deontological decision. Trust was greater if the vehicle made a moral decision aligned with the participants' values.

These prior studies show that people trust AI advice that aligns with their values more. In these studies, the AI's recommendations are seen as value-laden since they pertain to ethical situations. It could be that people also perceive inaccuracies in AI recommendations as value-laden judgments in such situations. This means that biases in AI advice might not just be seen as errors but also as advice that aligns with certain values. For example, a medical AI that always gives recommendations that are overly risk-averse might be perceived as cautious. The similarity between an AI's bias and a user's personal values might influence trust and adoption of the advice. If this is the case, calibration of AI recommendations to align more closely with user values might alleviate the loss of trust due to errors. Riedl (2022) and Mehrotra et al. (2021) have stated that adaptive systems that take user values and characteristics into account are essential to increase the adoption of decision-making aids in subjective scenarios. Furthermore, understanding the interaction between bias and trust in AI can provide insight into how existing system biases may result in over- or under-reliance depending on user beliefs. The primary aim of the present research is to study whether the similarity between user values and AI bias influences trust. This is formulated in the following research question:

RQ1: How does value similarity influence trust in biased AI advice?

Previous studies on AI value similarity utilize hypothetical decision-making scenarios that lack user input and engagement. Participants observe as the system makes a decision and are questioned on their trust in the system afterward. Some studies include self-reported willingness to act but none measure trust-related behavioral changes. Self-reported measures do not always translate into behavioral changes (Pharmer et al., 2021). Therefore, trust-related behavioral measures are collected in addition to self-reported trust measures in the present study.

Trust in AI advice develops over time (Cabiddu et al., 2022). Seeing an AI function repeatedly and being able to evaluate its outputs influences trust formation and behavior. Most value similarity studies do not incorporate repeated interactions with the AI. When they do, trust fluctuations are not measured (Yokoi & Nakayachi, 2019; Liu & Moore, 2022). Decision aids are used repeatedly in real-world applications. Thus it is crucial to understand the development of trust and compliance over time. Additionally, a user’s perception of biased advice may not occur after a single interaction when the AI's values are expressed subtly. For these reasons, a second research question is posed:

RQ2: How does value similarity influence trust in biased AI advice over time?

RQ1 and RQ2 assume that the user perceives value similarity with an AI. That is to say, users explain the behavior of the AI through social values that they can align or dis-align with. However, Mehrotra et al. (2021) found that manipulations of values do not necessarily lead to a difference in perceived value similarity. Gray et al. (2012) and Cvetkovich (2013) have argued that judgments of value similarity rely on a person's understanding of how the human mind works. For an entity to be perceived as a moral decision-maker, it must be perceived as having a mind. Shank et al. (2021) find that, under certain circumstances, people do attribute a mind to AI. However, whether users perceive values implied through AI behavior and whether this results in value similarity is unknown. Therefore, an additional exploratory research question is examined:

RQ3: How does mind perception influence value similarity?

The three research questions are investigated using a prison sentencing task. Participants are tasked with determining adequate prison sentences based on short descriptions of criminal cases. They receive advice from an AI (or human in the control condition). Bias is introduced to the AI by having it advise strictly lower or higher prison sentences than the final verdict of a judge. Value similarity is based on the participant's attitudes towards prison sentencing compared to the AI's bias.

The rest of this report is structured as follows: first, a literature overview discussing the relevant theories and prior works, resulting in the formulation of hypotheses. A detailed overview of the study design, study procedure, and measurement instruments follows this. Afterward, the analysis of the results is presented and discussed. To conclude, the limitations of the study, potential future studies, and implications of the results are discussed.

Theoretical background

Terms such as AI, algorithm, and system are used interchangeably in the literature on decision-making aids. This report will refer to these technologies as AI unless specifically referring to a technology discussed in prior work. The OECD (Organisation for Economic Co-operation and Development) definition of AI will be used: "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments…AI systems are designed to operate with varying levels of autonomy" (Lockey et al., 2021).

Trust

Definition of trust

Trust is defined in several ways in the AI literature. In AI-assisted decision-making, which is the focus of the present study, the most referred to definition is proposed by Lee & See (2004) (Vereschak, Bailly & Caramiaux, 2021). They define trust as “An attitude that an agent will achieve an individual's goal in a situation characterized by uncertainty and vulnerability.” This definition concisely captures the two key aspects of trust, uncertainty, and vulnerability. Uncertainty indicates that the realization of the individual's (the trustor's) goal by the agent (the trustee) is not guaranteed. Vulnerability indicates that the failure of the trustee to realize the trustor's goal will have negative consequences for the trustor. In addition to these two aspects, Vereschak et al. (2021) identify a third element that, while not explicitly stated in the above definition, is viewed as a defining element of trust, positive expectations. Lockey et al. (2021) state that, to speak of trust, the trustor must have positive expectations of the intention or the behavior of the trustee. In the case of AI, positive expectations are the expected utility or realized value from use. Uncertainty, vulnerability, and positive expectations are what constitute trust. These three aspects must be present for trust to be relevant in an interaction.

Psychological trust

In the above definition, trust is an attitude. Trust is a way of thinking or feeling about the trustee. This trusting attitude is a psychological construct. It exists in the mind of the trustor and cannot be directly observed or measured. This aspect of trust will be referred to as psychological trust. Prior research on trust measurements has identified various factors associated with trust (Schaefer et al., 2016). For example, peer recommendations, task expertise, and cultural background. The present study's interest is mainly in task-related trust factors. These factors influence the current task (Hoff & Bashir, 2015). For example, the AI's accuracy or how advice is presented to the user. Madsen & Gregor (2000) propose a human-computer trust model based on task-related factors. They propose five base constructs of computer trust: Understandability, technical competence, reliability, personal attachment, and faith. These constructs are further subdivided into cognition-based trust and affect-based trust. A visualization of their trust model is seen in Figure 1. In the current study, psychological trust will be assessed based on this model.

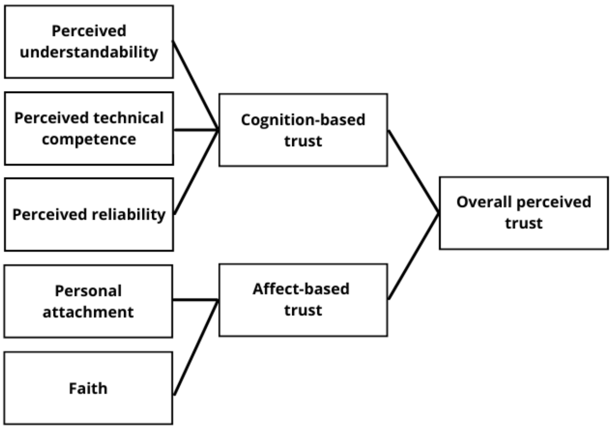


Figure 1: The human-computer trust model as proposed by Madsen & Gregor. Overall perceived trust is split into two separate aspects: Cognitive- and affect-based trust. These aspects are further divided into five constructs: Understandability, technical competence, reliability, personal attachment, and faith.

Behavioral trust

It seems logical to expect that users who report higher levels of trust would be more willing to comply with advice. While this is often the case (Riedl, 2022), psychological trust does not always result in a behavioral change (Pharmer et al., 2021). One explanation may be that a certain trust threshold must be reached before a behavior change occurs, especially for a binary choice (Vereschak et al., 2021). Chancey et al. (2015) suggest that multiple variables cause the dissociation between psychological measures of trust and compliance with AI recommendations. People may have different notions and interpretations of trust. Results from Liu & Moore (2022) support this idea. In their study, an AI recommends whether a professional decision-maker should be investigated for being politically biased. For example, a banker was suspected of refusing loans to same-sex couples. They found that alignment between the socio-political beliefs of participants and the recommendations of the AI positively influenced behavioral intention but not self-reported trust. They conclude that there is a difference between trust in AI and the acceptability of the advice. Someone that does not trust an AI can still agree with its advice and choose to act upon it.

Given the discrepancies between self-reported trust and behavior, Vereschak et al. (2021) recommend including trust-related behavioral measures alongside measures of psychological trust. Trust-related behavioral changes are changes in user behavior that result from trust formation. The most commonly measured trust-related behavioral change in prior literature is compliance with AI advice (Vereschak et al., 2021). Compliance refers to users following the AI's recommendation. Compliance can vary from a minor shift in a user's initial decision to a complete adoption of advice. Other behavioral measures exist. Feng & Boyd-Graber (2019) measure participants' decision-making times when presented with AI advice. A lower decision-making time (faster response) is associated with greater trust. For systems where advice is not automatically provided, reliance may be measured in addition to compliance. Yu et al. (2017) measure reliance based on how often a participant requests advice or delegates a decision to an assistive AI.

Factors influencing trust in AI

Trust in AI is influenced by numerous factors, including digital literacy (Logg, Minson & Moore, 2019), prior use by others (Alexander, Blinder & Zak, 2018), general trust (Höddinghaus, Sondern & Hertel, 2021), expertise (Logg, 2017), domain (Green & Chen, 2019), task complexity (Fan et al., 2020) and algorithm accuracy (Dietvorst et al., 2015). An extensive overview of factors influencing trust in AI can be found in the works of Hoff & Bashir (2015) and Schaefer et al. (2016).

One of the main factors that negatively impact trust is AI error. Dzindolet et al. (2003) found that an overall reliable decision aid that was initially viewed as trustworthy was distrusted after the aid made even a single error. Dietvorst et al. (2015) found a similar result. In their study, participants who witnessed an algorithm err preferred the decisions of a human decision-maker. This result remained even after participants observed a human make more mistakes than the algorithm. Dietvorst et al. (2015) refer to this preference for human advice as algorithm aversion. Many other studies find similar results (Burton, Stein & Jensen, 2020). Hoffman et al. (2018) explain the drop in trust after seeing an AI err as a swift reaction to surprise resulting from unrealistic expatiations. Renier et al. (2021) found that people are more forgiving toward humans than AI. They show that people expect AI is always perfect. In contrast, humans are expected to make occasional mistakes.

Due to the changes in trust resulting from interaction with AI, Hoffman et al. (2018) recommend treating trust as a dynamic process instead of a static quantity measured at the end of an experiment. In the present study, we are particularly interested in these influences of errors on trust development. For one because a real-world system would be used repeatedly, resulting in trust morphing over time. Thus the usefulness of improving only initial trust is questionable. And second, the AI advice in the current study will be manipulated to intentionally make errors that align or dis-align with the user's beliefs.

Appropriate trust in AI advice

Lack of trust is one of the main reasons for the low acceptance of AI decision aids (Dzindolet et al., 2003). The goal of trust interventions should, however, not be the arbitrary improvement of trust. Trust should promote appropriate reliance. Trust should be calibrated so that users trust reliable AI and distrust unreliable AI. Inappropriate reliance on AI can be separated into two categories: misuse and disuse (Parasuraman & Riley, 1997). Disuse is defined as the underutilization of automation. Disuse results in the potential benefits of AI not being fully realized. Misuse is defined as overreliance on automation. This behavior is commonly referred to as automation bias (Skitka, Mosier & Burdick, 1999). It is characterized by complacency and can be dangerous when advice is being relied on in critical situations, such as medical decisions.

Dzindolet et al. (2003) found that a trust intervention aimed at preventing misuse can lead to disuse. Their study observes a loss of trust in a decision-making aid due to the system erring. They find that explaining the system error improves trust, however, to such an extent that participants were willing to trust unreliable systems. This is of interest to the present study. An AI that purposefully makes recommendations that the user is more likely to agree with may be more trusted but simultaneously be more prone to misuse. Therefore, it is important to evaluate user trust relative to the AI's performance and carefully assess whether appropriate reliance is formed.

Value similarity

One of the earliest works reporting the influence of shared values on trust is the study by Sitkin & Roth (1993). They found value congruence to influence interpersonal trust positively. Earle & Cvetkovich (1995) proposed a general model of salient value similarity (SVS). This model suggests that a trustor and trustee sharing one or more salient values positively affects trust. For a value to be considered a 'salient value' it must be relevant to the current situation and important to the trustor. Furthermore, the trustor must be aware that the trustee holds this value. Attribution of values to the trustee is based on their actions, statements, and identity (Cvetkovich, 2013). For instance, seeing a person separate their garbage signals that they value sustainability. And reading a company's mission statement gives insight into the values that they consider important. This attribution of values is a fast and automatic heuristic process that often precedes other trust judgments (Poortinga & Pidgeon, 2006).

SVS theory has been shown to predict social trust in various situations (Siegrist, Cvetkovich & Roth, 2000). For example, Twyman, Harvey & Harries (2008) showed that value similarity improves trust in advice from government and consumer agencies. In their study, participants filled out questions about general values (e.g., money matters more than most things). Afterward, they were assigned an advisor and informed how the advisor had answered the same questions. They were then tasked with risk assessment of various activities (e.g., extreme sports, hazardous jobs, or drug use). Participants expressed greater trust in advice and were more compliant with advice if the advisor had given similar answers to the value questions. Gigliotti et al. (2020) applied the SVS framework in practice. They analyzed landowners' trust in the natural resources department of the government. The landowners were questioned on their values regarding wildlife management (e.g., are humans more important than animals). They found that landowners with a similar value orientation as the government trust government programs more.

The core aspect of the SVS theory is values, yet the definition of a value is inconsistent in the literature using SVS. Lee & See (2004) found that shared fundamental cultural values positively influence interpersonal trust. Lee & See (2004) describe the sharing of cultural values as people obeying the same rules and assumptions. Values can be defined as desirable goals or beliefs of varying importance that serve as guiding principles in people's lives (Schwartz, 1994). In other literature, values are viewed as the fundamental building blocks of beliefs and goals (Ives & Kendal, 2014). These definitions clearly show that values are closely related to beliefs and goals. Given that it is difficult to distill a specific situation down to core values (e.g., autonomy, health, or achievement) it may be more practical to measure an overarching belief or goal instead. An example of this is the work of Cazier, Shao & Louis (2006). In their study value similarity is operationalized by informing participants that their moral and political beliefs are (dis)similar to an E-business. They make no mention of the specific values underlying those beliefs. Their results are in line with SVS theory, participants' trust is greater in businesses that share their beliefs. The present study will use a similar approach. Value-similarity will be examined based on the participants' beliefs. In the prison sentencing task, the most prominent belief is punitiveness (Adriaenssen & Aertsen, 2015), a person's belief of how strictly offenders should be punished.

Value similarity between humans and AI

If value similarity improves the trusting attitude towards advice in human-human relations, these effects might also occur in human-AI interactions. Riedl (2022) has argued that adaptive systems that take user personality traits and characteristics into account are essential to increase trust in and adoption of such systems. Mehrotra et al. (2021) further highlight the need for taking the value-based reasoning of users into account when developing decision-making aids.

A limited number of studies have looked at the influence of value similarity on trust in the case of human-AI interactions. Yokoi et al. (2020) found that participants perceive value-similarity between themselves and a medical AI prescribing treatment. In a series of follow-up studies, Yokoi & Nakayachi found shared moral beliefs to positively influence self-reported trust in autonomous cars (Yokoi & Nakayachi, 2021a) and a medical emergency decision-making system (Yokoi & Nakayachi, 2021b). In their studies, value similarity is based on users making the same ethical decisions as an AI. First, participants are presented with a scenario (e.g., Trolley problem) and asked to make a choice. After this, they are shown how the AI acted in the same scenario. Participants expressed greater trust in AI that made the same decision as them. Similar results were found by Mehrotra et al.

(2021). In their study participants in a hostage extraction scenario interacted with multiple different AI agents. Participants' value profiles were determined at the start of the study based on questions about core values (e.g., power, hedonism, conformity, etc.). They then consecutively received advice from five AIs with distinct value profiles. These value profiles were expressed through how the AI handled the hostage situation. For example, does it permit harming the hostage-takers? Participants reported higher levels of trust in an artificial agent with a similar value profile to them. Not all human-AI value similarity studies find positive results. Yokoi & Nakayachi (2019) found that value similarity with an AI in an investment game did not affect psychological and behavioral trust. They speculate that this is because the participants' main goal was to score well during the game. Because of this, their values regarding risk-taking were perhaps less relevant.

Similarly to the value similarity literature in human-human relations, different ways of referring to values are used in human-AI studies. Verberne, Ham & Midden (2012) found shared driving goals to increase trust in an adaptive cruise control system. And Liu & Moore (2022) found that political belief alignment increases willingness to act based on AI recommendations in numerous politically charged scenarios. This again highlights the close conceptual connection between values, goals, and beliefs.

A common denominator in studies analyzing the relation between value similarity and trust in AI is that experiments are based on scenarios. Participants are presented with a situation in which a machine makes a value-laden decision or gives value-laden advice. Participants do not directly interact with the AI. Instead, they read or observe. Psychological trust and behavioral intentions are determined solely through self-reported measures. The system's values, beliefs, or goals are explicitly presented to the participant in these studies. In the present study, we investigate whether the positive effects of value similarity on self-reported trust translate into behavioral changes. Furthermore, we examine whether these effects remain when system values are not explicitly presented to a participant but rather implied through AI behavior.

Value expression through AI bias

One way in which values may be implied through AI behavior is by systematically deviating the AI's output in favor of a specific belief. This way, a bias is purposely introduced to the AI. Here we are focusing on bias in a social context, as opposed to the statistical definition of bias. Bias, in this context, refers to the systematic deviation of AI outcomes resulting in unfair decisions (Mehrabi et al., 2021). For example, by favoring one group of users over another based on arbitrary inherent or acquired characteristics (Ala-Pietilä et al., 2020). AI systems sometimes reflect and perpetuate human social biases (Kordzadeh & Ghasemaghaei, 2022). This can, for instance, be the result of beliefs held by the AI's developers or biased training data. Biases in AI can perpetuate existing stereotypes and prejudices, leading to discrimination in decision-making tasks. For example, an algorithm may be less likely to approve immigrant loan applications (Mehrabi et al., 2021).

In the case of decision-making or advisory AI, the source of bias could be interpreted in different ways by users. Bias might be seen as the result of an error in the computations performed by the system. A technological flaw. On the other hand, bias could be perceived as a value-laden judgment by the AI itself. One that a user could be more willing to accept if their beliefs on the subject align with the AI's recommendation (Liu & Moore, 2022). Especially when the recommendation is related to a topic that is perceived as subjective and reliant on moral judgments, such as prison sentencing, loan approval, or personnel hiring. However, whether value similarity results from users perceiving AI systems as possessing values is unclear. Tolmeijer et al. (2022) found that users attribute faults of

AI not to the system itself but to its developers. The same might be true for values. Users may view biased or value-laden advice from an AI as a representation of the beliefs of its developers.

As previously discussed, designing AI systems for appropriate reliance is essential. The present research examines whether value-similarity can be used to increase compliance in AI. This provides insights into adaptive system design based on user values which may improve interactions between systems and users. However, it can also provide insight as to how existing system biases may result in over- or under-reliance depending on user beliefs and bring additional attention to the possible dangers of biased AI, especially in critical moral decision-making situations. In the present study, an AI expresses bias by systematically advising lower or higher prison sentences than the judge. The expectation is that when this bias aligns with a person's own beliefs on punitive measures, this results in value similarity.

Mind perception

Prior literature does not examine the source of value similarity in human-AI interactions. Neither do they assess if participants view an AI as capable of holding or acting based on values, goals, or beliefs. Cvetkovich (2013) argues that judgments of value similarity rely on a person's understanding of how the human mind works. This raises the question of whether value similarity between a person and an AI is possible in the first place.

An AI does not possess a human mind. As of yet, even the most sophisticated machine-learning systems cannot emulate the human mind. According to Gray, Young & Waytz (2012), mind perception is the foundation of moral judgment. For an entity to be perceived as a moral decision-maker, it must be perceived as having a mind. Thus, for value similarity to exist similarly to human-human relations, the user must perceive the AI to have some form of mind. Users must perceive AI to be capable of holding and/or acting based on values. In the mind perception literature, a distinction is made between two aspects of the mind, the agentic and experiential mind. The experiential mind refers to an entity's capacity to experience sensations. The agentic mind refers to an entity's capacity to act independently. If an agent is perceived to possess an agentic mind, their actions may be viewed as moral decisions. Gray, Gray & Wegner (2007) found that people perceive machines to have a mind. However, only the agentic aspect. Shank et al. (2021) found that mind perception, both agentic and experiential, influences the virtuous character attribution of AI and humans. AI systems which were perceived to have a greater capability of thought were more likely to be explained based on social factors. This is opposed to typical evaluations one may expect of a computer, e.g., based on efficiency or robustness.

Alternatively, even if AI is not consciously perceived as a moral actor, values and value similarity may still play a role in trust. The media equation hypothesis, proposed by Reeves and Nass (1996) found that people respond socially to technologies. This suggests that people may (subconsciously) perceive the behavior of an AI as being based on human-like social factors. Given that in human-human relations, values play a substantial role in forming trust. This may thus also be the case for human-AI relations. The actions and behavior of an AI might be viewed as an expression of its goals, values, or beliefs, similar to dispositional character attributions in human-human relations. An additional factor that could strengthen this effect is the black-box nature of complex AI systems. An increase in the complexity of AI has led to a decrease in technical understandability by users. This may drive people to be more inclined to explain the outputs of these systems using familiar mental models. An example of this is the work of Pauketat & Anthis, 2022. They found that participants attribute more emotional capacity to complex future AI than to simpler systems. This influence of system complexity may explain why Gray et al. (2007) did not find users to perceive a mind in a relatively simple machine, while Shank et al. (2021) did find this for their more sophisticated AI.

Present study

The present study aims to investigate whether the positive influence of value similarity on self-reported trust in AI advice translates to behavioral changes. Previous studies have shown that value similarity increases psychological trust. Additionally, some studies find behavioral intention to be positively impacted by value similarity. In human-human relations, the influence of value similarity on psychological trust is strongly correlated with a willingness to comply with advice and accept decisions. Therefore, psychological and behavioral trust is expected to be positively affected by value similarity with an AI advisor. This expectation is formulated in H1, which is the main hypothesis of interest in the current study.

H1: Value similarity increases (i) psychological and (ii) behavioral trust in an AI advisor.

An additional control condition in the form of a human advisor is tested. This aims to confirm that the results from previous value similarity studies are repeated in our setup. This is to ensure that we have successfully designed an experiment in which participants perceive values as a relevant aspect of the decision-making process, that the user perceives the values expressed by the advisors, and that value similarity can form between the user and advisor. Based on effect sizes reported in prior research, we expect that the influence of value similarity is stronger in human-human interactions.

H2: The effect of value similarity on (i) behavioral and (ii) psychological trust is greater for a human advisor than for an AI advisor.

Trust in AI is lower for tasks that require subjective decision-making, as opposed to mechanical or computational tasks (Castelo, Bos & Lehmann, 2019). Fair prison sentencing is seen as a subjective task. Therefore, we expect both psychological and behavioral trust in advice to be overall larger for human advisors.

H3: (i) Psychological and (ii) behavioral trust in advice is greater for a human advisor than for an AI advisor.

These three hypotheses are visualized in Figure 2.

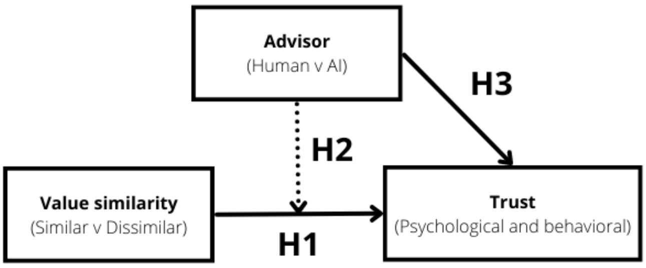


Figure 2. Model of the expected direct effects and interaction of value similarity and advisor type on psychological and behavioral trust.

Prior studies found that AI errors have a negative influence on user trust. Allowing users to interact with a system repeatedly and showing them the system is not perfect and makes occasional mistakes negatively impacts trust. This results in trust decreasing over time. Specifically, there seems to be a considerable difference between initial trust and dynamic trust. In the present study, the AI's mistakes will either align or dis-align with a user's preexisting beliefs. It is expected that value similarity moderates the negative effect of use over time, with the effect being weaker if values are similar and stronger if values are dissimilar.

H4: The effect of use over time on (i) psychological and (ii) behavioral trust in AI is moderated by value similarity, with the effect being weaker for greater value similarity.

In contrast to prior studies, the AI's values, beliefs, or goals will not be explicitly mentioned but are instead implied through biased advice. An additional exploratory part of this study is to examine if users perceive values being implied in such a way and whether this leads to value similarity. Gray et al. (2012) and Cvetkovich (2013) argued that some degree of mind perception must be present for value similarity to be relevant in trust formation. Suppose users do indeed perceive the AI to possess an (agentic) mind. In that case, it is expected that the degree of mind perception positively influences value similarity with the AI. It could be the case that no such relationship is found, even when value similarity is shown to influence trust and compliance. Such a finding would suggest that value similarity does not stem from people's perception of the inherent values of AI. Instead, it would provide evidence for the theory that people view AI's advice as a representation of the values and beliefs of its developers.

Study design

The present study consists of an experiment with a 2x2 factorial design. Advice and advisor are the factors, and psychological trust and compliance are the measured outcomes. Participants are tasked with determining prison sentences for various crimes based on short case descriptions. After determining an initial sentence they receive advice from an advisor. They are then given the option to change their initial sentence. Afterward, the participants are shown the prison sentence that the judge handed out in the real-world case. This task is repeated for 14 different cases.

Advice is manipulated to be lenient or strict. In the lenient condition, the advice provided to participants is always lower or equal to the real-world sentence handed out by the court. The reverse is true for the strict condition. Advice is randomly generated during the experiment and varies for each participant and case. Advice has a maximum deviation of 20% above or below the real-world sentence. Within this 20% deviation range, advice is uniformly distributed, resulting in an average deviation of 10%. The advice can be the same as that of the judge (0% deviation). Advice is provided to the participants as a recommended prison sentence of months of jail time. Participants are randomly allocated to either condition with an even distribution.

The advisor is manipulated to be an AI or a human. The human condition is present as a control group to verify that the designed scenario has relevant salient values that play a role in trust formation. In the AI condition, the advisor is referred to as 'legal AI' throughout the study. In the human condition, the advisor is instead referred to as a 'legal expert'. Besides this, no other changes to wording or phrasing are made throughout the study, both in the descriptions and in the measurement instruments. There is one exception to this, the first time the legal AI is introduced, an additional explanation stating that the legal AI is an "artificial intelligence decision support system" is added. This is done to ensure participants are aware of what AI stands for and realize that this is not a human. Participants are randomly allocated to either condition with a 70% chance of being assigned to the AI condition. The primary interest of this study is the effect of value similarity on trust in AI.

Material development

The experiment is created using lab.js. Lab.js is a JavaScript-based free, open-access, online experiment-building tool developed by Felix Henninger (Henninger et al., 2022). The builder allows for the randomization of variables and the assignment of participants to (weighted) random conditions. This made the tool suitable for the generation of individually randomized advice for each participant. All experiment material (introduction, task instructions, advice provided, and questions) is adjusted to allow AI and human conditions to be used interchangeably as a way of referring to the advisor.

During the experiment participants read descriptions of committed crimes and were tasked with determining appropriate prison sentences. The descriptions consist of summarized versions of case files from real crimes that recently occurred in The Netherlands. Case files were retrieved from rechtspraak.nl. This is an open-access database that publishes and maintains an archive of verdicts in the Dutch legal system. Cases were selected based on the following criteria: sentencing date, sentence duration, type of offense, and case complexity.

Sentencing date

To ensure consistency between verdicts and account for changes in legal procedures over time, case verdicts from a specific time frame (2020-2022) were taken. The potential problem of participants being familiar with the cases due to their recency (e.g., case featured in the news) is negated by participants not being Dutch citizens.

Sentence duration

The case verdict by the court must be at least 16 months of jail time. This minimum is set to ensure that the deviation of the advice from the actual jail sentence results in a clear difference. For instance, in a case where the jail sentence handed out is seven months, a maximum deviation of only a single month would be possible (six months in the lenient condition and eight months in the strict condition). This would result in the advisor often giving the same sentence as the judge, resulting in no perceivable bias in the advice.

Type of offense

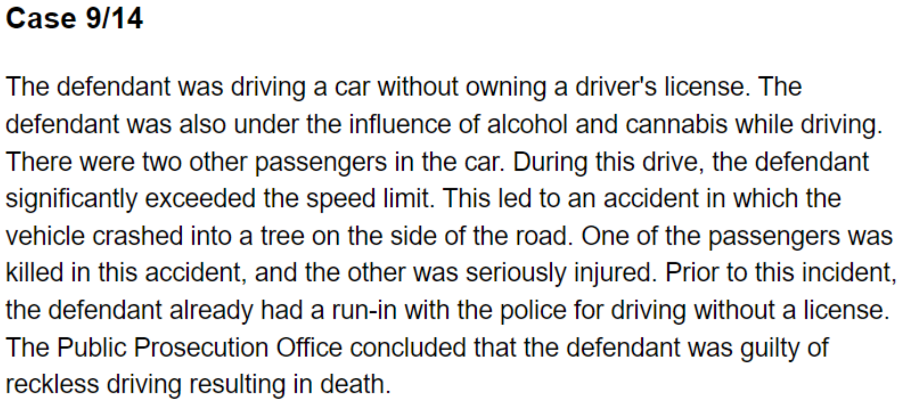
The type of offense must meet three criteria. First, the offense must have been a criminal offense committed towards another person and have been trialed as such by a criminal court. This means that offenses such as tax evasion are not considered. Second, only cases in which the verdict mainly consisted of a prison sentence were selected. Participants are only allowed to judge the cases by estimating a prison sentence, and no additional penalties, such as fines, rehabilitation, etc., can be handed out by the participants. Cases that include such additional punitive measures would likely be confusing to participants since the court's sentencing takes these additional measures into account, whereas the participants cannot. Finally, the offense may not be primarily or in part related to sexual misconduct such as sexual harassment, sexual assault, or rape. Similarly, the offense may not involve, in any way, children under the age of 18. These restrictions are set to ensure that participants are not confronted with especially disturbing cases.

Case complexity

Case complexity is based on the number of separate crimes that were committed and taken into account in the judge's verdict. Separate crimes of different natures are seen as more complex, e.g., a combination of murder and robbery is more complex than two robberies. Additionally, possibly alleviating and extenuating circumstances surrounding the crime were viewed as increasing case complexity. Examples of alleviating circumstances are the perpetrator's young age, bad social background, or substance addiction. Examples of extenuating circumstances are repeated offenses, lack of remorse, or 'professional' criminality. These additional factors were only considered to contribute to case complexity if it was specifically mentioned in the original case file that the court considered them in their verdict. The choice was made to focus on cases with relatively low complexity to ensure participants, who were expected to have a low level of legal expertise, could still make somewhat accurate decisions and were not completely reliant on the advisor. Previous research has shown that when a task becomes, more difficult, people are more likely to rely on advice (Fan et al., 2020). This may overshadow the trust and compliance resulting from value similarity.

In addition to these four selection criteria, an effort was made to ensure a broad spread of offenses. This is done since similar crimes tend to get similar sentences in The Netherlands. For instance, if all cases had been related to drug trafficking, participants would likely realize that the prison sentence would always be around 24 months. In addition to this, a spread of offenses gives a clear view of a participant's different interpretations of various crimes, which can more easily be related to the measure of general punitiveness.

Based on the above process, a total of 14 legal cases were selected. Crimes include murder, assault, drunk driving, (armed) robbery, (online) scamming, prison break, extortion, and drug trafficking. Prison sentence duration ranged from 16 to 60 months. The descriptions were created by summarizing the original case files into the most important details of the crimes committed. The important factors were mainly determined based on which factors the court mentioned in their final verdict. In the summarized descriptions, all people are referred to using gender-neutral pronouns (they/them/their). This is done because the original case files make no consistent mention of gender. The criminal/defendant/suspect/perpetrator is always referred to as 'the defendant' for the sake of consistency. The cases will be read by UK citizens who will be told these crimes occurred in the UK. Therefore, any references to The Netherlands have been removed, and monetary quantities are expressed in pounds instead of euros. We chose to tell participants the crimes occurred and were trialed in the UK for two reasons. Firstly, we wanted to avoid people speculating on how a prison system in a foreign country functions. Second, it is expected that people's opinions on punishments may differ when crimes are being committed in their home country as opposed to abroad. At the end of each case description, the verdict of the 'Public Prosecution Office' is given, this serves as a short one-sentence summary of the case. An example case description, as it was presented to the participants, is shown in Figure 3.



Case 9/14

The defendant was driving a car without owning a driver's license. The defendant was also under the influence of alcohol and cannabis while driving. There were two other passengers in the car. During this drive, the defendant significantly exceeded the speed limit. This led to an accident in which the vehicle crashed into a tree on the side of the road. One of the passengers was killed in this accident, and the other was seriously injured. Prior to this incident, the defendant already had a run-in with the police for driving without a license. The Public Prosecution Office concluded that the defendant was guilty of reckless driving resulting in death.

Figure 3: Example of a case description as it was presented to participants during the study.

Measures

Trust and value similarity are measured in multiple ways. Verified scales as well as matching between beliefs and advice conditions are used. Additionally, personal beliefs and demographic data are collected.

Psychological trust

Psychological trust (self-reported trust) in advice is measured using an 11-item scale adapted from Madsen & Gregor (2000). This measurement instrument was chosen as it has been developed, tested, and verified specifically for measuring trust in automation technologies. The original scale consists of 25 items. This scale identifies and measures five separate sub-factors of trust: reliability, technical competence, understandability, faith, and personal attachment. For the present research, the sub-scales on understandability and personal attachment were not included. The understandability sub-scale relates to the ease of use of the system. The personal attachment sub-scale measures participants' attitudes toward anticipated long-term use of the system. Neither of these aspects is relevant to the current research setup.

In addition to excluding the two subscales, four items are dropped. Two because of redundancy (Hoffman et al., 2018) and the aim to keep the questionnaire short. And two for not applying to the human advisor condition. For some individual questions, small changes to wording are made to ensure the questions are usable in both the AI and human condition, e.g., removal or replacement of the words 'it' and 'the system' in reference to the advisor. Madsen & Gregor (2000) do not provide information on how they administered the scale in their original research. Similar to other works that adapt this scale, responses will be measured using a Likert scale with seven answer options ranging from strongly disagree to agree strongly.

In addition to this scale, which participants will answer after completing all cases, three short and intuitive trust-related questions are asked after every case. These questions aim to measure the change in psychological trust over time. The questions are adapted from the Madsen & Gregor faith sub-scale of trust and have been modified to be shorter and easier to understand. This is done to ease the load on the participants as they are required to answer these questions 14 times. Answer options range from 1 - strongly disagree to 7 - strongly agree.

Behavioral trust through weight on advice

To analyze trust-related behavioral changes, compliance with the recommendations is measured. Compliance is measured through weight on advice (WOA) (Harvey & Fischer, 1997). WOA is a commonly used measure of compliance or advice utilization in judge advisory tasks (Himmelstein, 2022). It measures the proportional change in a decision after advice is received. The definition of WOA is as follows:

$$WOA = \frac{\text{initial decision} - \text{final decision}}{\text{initial decision} - \text{advice}}$$

In the current study, the initial decision refers to the prison sentence determined by the participant before receiving advice. The final decision is their (possibly adjusted) sentence after having received advice. The advice is the jail sentence recommended by the advisor. All values are expressed in months of jail time. Assuming rational decision-making by the participants, WOA should result in a measure ranging from 0 (no advice adoption) to 1 (complete advice adoption). It is, however, possible to have a WOA outside of this range. For this reason, research utilizing the WOA measure tends to set 0 and 1 as boundary values. There is no clear consensus in the literature on how to handle values outside this range (Bonaccio & Dalal, 2006). Prahl & Van Swol (2017) opted to include values outside the range in their analysis. They interpret negative values as a strong aversion to advice. Logg, Minson & Moore (2019) instead choose to winsorize the WOA measure. All values outside the boundary range are rounded toward their closest boundary value. The problem with this method is that overshooting advice is then interpreted as complete adoption. Some authors instead use a version of WOA with absolute difference terms (Yaniv, 2004). This results in ambiguity between compliance and rejection of advice and is therefore not suitable for the present study. Himmelstein (2022) argues that, regardless of interpretation, values outside the 0 - 1 range should be excluded from statistical analysis as they violate the continuous nature of the measure. Bonaccio & Dalal (2006) suggest discarding values outside the boundary range as long as most data is within it. This will be the approach in the present study.

WOA values outside the 0-1 range will be investigated separately as they may be indicative of participants being aware of the experimental manipulation. For instance, a participant who realizes that the advice they are receiving is biased towards higher sentences may choose to lower their sentences to be below that of the advisor, irrespective of their initial sentence. This would result in their sentences being as close to the judge's decisions as possible. They might choose this 'strategy' if they have no particular opinions on sentencing durations and would rather keep their sentences in line with the judge's opinion. In an attempt to prevent this type of gamification of the task, participants are instructed multiple times to hand out sentences that they would feel comfortable giving the defendant in the real world. They are told to view the decision by the judge as a reference value instead of the correct answer. After all tasks are completed, an additional open question is posed to the participants, asking them whether they felt they were able to envision their sentences being handed out to defendants in the real world.

Self-reported value similarity

Value similarity between the participant and the advisor is determined using two methods. Value-similarity is assessed directly using the value similarity scale from (Siegrist et al., 2000). This scale consists of five semantic differentials (e.g., Same values – Different values). Six response categories are given, ranging from one to six. The original scale uses seven response categories which allow for a neutral answer. The decision to use an even number instead was made to force participants to choose, to encourage them to consider their answer carefully, and to eliminate problems with interpreting the mid-point answer (what does it mean to be neither similar nor dissimilar to someone?).

Value similarity based on punitiveness and experimental condition

Salient value similarity theory suggests that value congruence with an advisor stems from the most salient value in the decision-making scenario. In the prison sentencing task, the salient values are related to the punitiveness of the participant. The second value similarity measure is determined by a participant's punitiveness and linking it to their assigned experimental condition (lenient or strict). This variable will be referred to as 'punitiveness by condition'. This measure of value similarity is expressed as a dichotomous variable that has the categories: value-similar and value-dissimilar. For example, participants who report high punitiveness and are assigned to the strict advisor condition are placed in the value-similar category. The distinction between high and low punitiveness is made based on the participant's answers to the punitive attitudes scale. A mean score greater or equal to four is considered high punitiveness, and a mean score lower than four is considered low punitiveness. A full overview of this value similarity variable is shown in Table 1.

Value similarity		
	Punitive attitudes	
Advice condition	High	Low
Lenient	Dissimilar	Similar
Strict	Similar	Dissimilar

Table 1: Value-similarity interpretation based on experimental advice condition (lenient, strict) and self-reported punitive attitudes (high, low). Value-similarity is split into two groups: similar and dissimilar.

Creating separate categories of value-similar and value-dissimilar in this way is based on previous value-similarity literature. Doing this allows for comparisons between groups. However, turning the continuous variable of punitiveness into a dichotomous variable by making an arbitrary division of high and low punitiveness at the scale's midpoint results in a lot of information being lost. This method will be used as an initial test to see whether our results align with previous findings. Afterward, value similarity based on punitiveness and advice condition will be analyzed through an interaction term in a regression to keep the continuous variable intact.

Punitiveness

Punitiveness (or punitive attitude) is the degree to which a person favors strict or harsh punishments for criminal offenders (Adriaenssen & Aertsen, 2015). The level of punitiveness is measured using a 7-item scale adapted from (Spiranovic, Roberts & Indermaur, 2012). Example item: *People who break the law should be given stiffer sentences*. All items are measured on a seven-point Likert scale ranging from 1 - Strongly disagree to 7 - Strongly agree. This scale is chosen as it predominantly measures a general level of punitiveness, in contrast to scales with a narrower focus which measure punitiveness for a single type of offense (e.g., murder). Furthermore, this scale solely focuses on prison sentencing. It does not include questions related to other forms of punitiveness, such as fines, community service, house arrest, etc. Prison sentencing is the focus of the current study and the only way through which participants can express their punitiveness, therefore, a scale that measures this particular aspect is preferred.

Mind perception

Mind perception is measured through a scale adapted from Li et al. (2022). This scale measures explicit attribution of mind perception to an agent. The original 15-item scale measures two distinct aspects of mind perception: agency and experience. As explained previously, in the current study, the interest is in perceived agency. Therefore the experience-related items are dropped. This leaves an eight-item scale measuring agency. Example item: *The legal AI can tell right from wrong*. Each item is measured on a seven-point Likert scale ranging from 1 - Strongly disagree to 7 - Strongly agree. This scale is particularly suitable for the present research as it has been developed to measure mind perception in both humans and machines. Because of this, questions did not need to be adapted to fit the separate conditions.

Procedure

The study was made available online to eligible participants through Prolific. Before deciding to participate in the study, participants had the option of reading a short description stating the topic and goal of the study. Upon deciding to participate, participants were greeted by the welcome screen and asked to read and fill out the consent form. After they gave their consent, they were asked to fill in their Prolific ID.

Next, they received a detailed introduction to the task and were introduced to the advisor (AI or human). Afterward, they completed one practice case. The practice case was identical to the real task but included additional explanations on each step of the procedure. The practice case was immediately followed by the actual task consisting of the 14 case descriptions. For each description, participants were requested to read the case and provide a prison sentence that they deemed reasonable. Participants gave their sentences in months of jail time (e.g., 2,5 years = 30 months). After submitting their sentence, they received advice from the advisor and were given the option to alter their initial sentence based on this advice. Finally, after each case, the participants were simultaneously shown their final estimate, the advisor's estimate, and the prison sentence that the judge handed out in the real-world case.

After completing all cases, participants filled out the value similarity, punitiveness, trust, and mind perception questionnaires. Additionally, participants answered a manipulation check and were asked how realistic and engaged they were in the task. To conclude the experiment, demographic variables were collected, after which participants were thanked for their participation and given the completion code (required for Prolific survey completion). The average completion time of the experiment was 19 minutes. Participants received compensation of £3,00 after review and approval of their response.

Participants

Participants were recruited through the online research platform Prolific. Participants were filtered by age (minimum age of 18), English language fluency, country (UK), political affiliation, and participation in prior similar studies. Political affiliation was based on a single question: "What political position aligns most with your beliefs" with answer options left, right, moderate, or skip. Only participants who self-identified as either left or right were selected. Moderate participants were excluded to increase the likelihood of participants having a clear stance on punitive measures. Political affiliation was determined to be the most suitable filter available in Prolific for this. The expectation is that left-leaning participants have lower punitive attitudes than right-leaning participants.

To ensure participants can easily understand the formal language used in the case descriptions, participants are required to be fluent in English. A large enough sample of English-speaking Dutch participants meeting all other requirements was unavailable. Therefore, UK participants were selected as the UK has a similar legal system to that of the Netherlands, meaning participants would be more familiar with Dutch prison sentence durations. Participants from the USA were excluded for this reason, with American prison sentences being considerably longer than those in the Netherlands.

This study was run alongside another study that used the same case descriptions. Additionally, Prolific had previously been used to conduct a similar experiment. Participation in any of these studies was filtered to avoid participation in multiple studies.

The recruited number of participants is based on a priori sample size calculation using G*power version 3.1.9.7. An estimation of effect size was made based on previous value similarity studies, which included some form of machine trust measure. This resulted in a sample size weighted mean effect size of $\eta_p^2 = 0.0659$. For studies that did not report effect size as η_p^2, the effect size was converted using the tools provided by Lenhard & Lenhard (2016). For alpha = 0.05 and a power of 90%, a sample size of 151 is required. Ten additional participants were recruited for a pilot study. The primary aim of the pilot was to ensure correct random assignment to condition and randomization of AI advice. No problems were found in the pilot, and no subsequent changes to the study were made. Therefore, the pilot participants were included in the data analysis. Due to minor irregularities with Prolific, the total number of participants in both studies combined was 171. Two participants were excluded from the analysis for failing to answer both attention checks correctly.

The final number of participants included in the analysis is 169 (87 female, 80 male, one other, and one prefer not to say). Further sample descriptive statistics can be found in table Table 2. Given the low legal expertise of the sample (M = 2.1), legal expertise was excluded from the analysis.

	Mean	SD	Median	Min	Max	Range
Punitiveness	4.79	1.47	5.14	1	7	1 - 7
Political affiliation	5.21	2.66	5	1	10	1 - 10
Legal expertise	2.1	1.68	1	1	8	1 - 10
Age	46.24	15.74	45	19	82	NA
Education	3.48	1.2	4	1	6	1 - 6

Table 2. Sample descriptive statistics. Education answer options were: Elementary school, high school, trade school, bachelor, master, and Ph.D.

Results

Data analysis and visualization are conducted in R. First, manipulation and assumption checks are performed. This is followed by testing the main hypotheses. To conclude, other potentially interesting findings in the data are presented.

Manipulation check

To ensure participants perceived the bias in the advice, they were asked to rate the advisor's punitiveness from 1 – lenient (short duration of prison sentences) to 10 – strict (long duration of prison sentences). Results are visualized in Figure 4. A 2x2 ANOVA with advice (lenient, strict) and advisor (AI, human) was conducted. Results show that perceived punitiveness is greater in the strict condition (M = 5.12, SD = 2.11) than in the lenient condition (M = 3.87, SD = 1.79) (F(1,165) = 16.45, p < 0.001, η2 = 0.09). This indicates that the manipulation was successful. Additionally, the AI advisor is perceived as being more punitive (M = 4.82, SD = 2.10) than the human advisor (M = 4.10, SD = 1.94) (F(1,165) = 4.42, p = 0.037, η2 = 0.03).

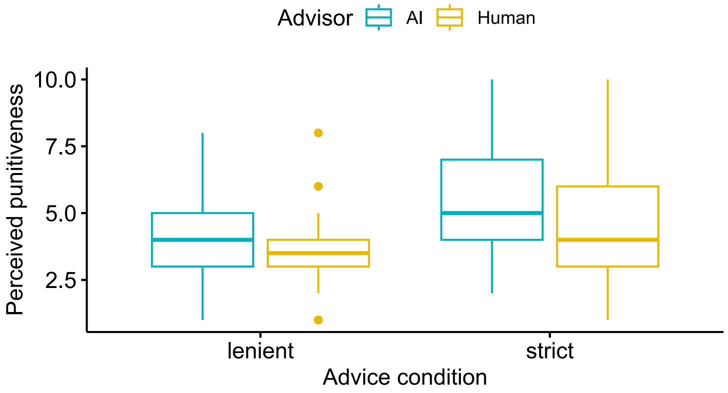


Figure 4. Boxplots showing the difference in perceived punitiveness between the advice and advisor conditions.

The relation between perceived advisor punitiveness and punitive attitudes was examined to verify whether participants are comparing the punitiveness of their advisor to their personal beliefs and not only to the decision made by the judge. While controlling for experimental conditions, punitive attitudes negatively predict the perceived punitiveness of the advisor (β = -0.16, SE = 0.08, p = 0.03). This shows that participants with greater punitive attitudes viewed the advisor as more lenient, irrespective of advice condition. This supports the expectation that people compare the advice to their beliefs.

Punitive attitudes distribution

Participants were recruited based on having a clear political preference (left or right). A broad range of punitive attitudes in the sample is required for the value similarity analysis. It was assumed that political affiliation is correlated with punitive attitudes. Regression analysis with punitive attitudes as the DV and political affiliation as the IV was performed to check this assumption. Age, gender, and education were included as covariates. To satisfy the regression assumptions, a square transformation is applied to the DV.

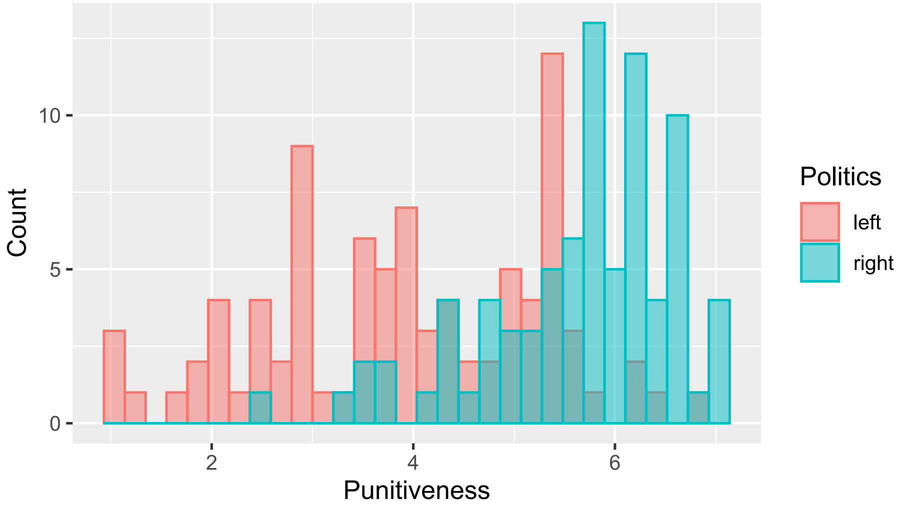


Figure 5. Participant punitiveness distribution in the sample split into left and right political affiliation.

Results indicate that there is a significant relationship between political affiliation and punitiveness (β = 0.55, SE = 0.073, p < 0.001). Participants with more right-leaning political beliefs have a stronger punitive attitude. This result supports our assumption. Results further show a negative relation between the level of education and punitiveness (β = -0.17, SE = 0.065, p = 0.01). Participants who completed a higher education degree were less punitive. None of the other covariates significantly impacted punitiveness. Figure 5 shows the punitiveness distribution among the sample based on general political affiliation (left or right). While the sample is, on average more punitive than expected, there is a broad enough spread to continue with the value similarity analysis.

Descriptive statistics and correlations

Table 3 shows descriptive statistics of all the measured variables. A distinction is made between overall trust measured at the end of the experiment using the 11-item scale (Final trust) and the combined average of the 3-item scale measured after each task (Case trust). The variable 'difference' refers to the difference between a participant's initial sentencing decision and the advice they received. 'Decision time' is the time (in seconds) that a participant took to make their decision after receiving advice. Given the low legal expertise of the sample (M = 2.1 on a 10-point scale), legal expertise is excluded from the analysis.

	AI (n = 118)				**Human** (n = 51)			
	Lenient (n = 48)		**Strict** (n = 70)		**Lenient** (n = 22)		**Strict** (n = 29)	
Variable	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range	Mean (SD)	Range
Final trust	4.06 (1.28)	1-6.5	4.59 (1.28)	1-6.5	4.75 (1.26)	1.1-6.7	4.76 (1.47)	1-6.8
WOA	.308 (.204)	0-.89	.456 (.220)	0-.98	.447 (.245)	0-1	.433 (.220)	0-.86
Value similarity	2.90 (1.12)	1-6	3.30 (1.25)	1-6	3.39 (1.13)	1-5.4	3.33 (1.22)	1-5.4
Mind perception	2.67 (1.18)	1-5.4	2.82 (1.24)	1-6	5.26 (.888)	2.8-6.8	5.33 (.756)	3.9-6.6
Case trust	4.10 (1.40)	1-6.8	4.54 (1.59)	1-6.9	4.59 (1.36)	1.24-7	4.85 (1.48)	1-7
Punitiveness	4.81 (1.42)	1-7	4.59 (1.49)	1-6.9	4.64 (1.18)	2.7-6.6	5.18 (1.66)	1.1-7
Difference	23.0 (12.5)	11-67	28.9 (29.9)	11-193	22.2 (15.7)	10-78	28.8 (19.6)	11-97
Decision time	8.05 (3.28)	3-17	8.82 (3.49)	4-20	8.05 (3.05)	4-14	9.95 (3.51)	4-17

Table 3. Descriptive statistics of all measured variables for each condition.

Table 4 depicts Pearson correlation coefficients of all measured variables in the AI condition (upper diagonal) and the human condition (lower diagonal). 'Punit by condition' refers to the dichotomous variable 'punitiveness by condition' constructed from participants' punitiveness and advice condition. Its values are 'similar' (baseline) and 'dissimilar'.

Both measures of psychological trust (Final trust and case trust) are strongly correlated. This is expected as they are adapted from the same scale. Both measures of psychological trust strongly correlate with WOA. There are moderate to strong correlations between all trust measures and value similarity. Difference in advice negatively correlates with trust, WOA, and value similarity. These correlations are considerably stronger in the human condition. Notably, the constructed variable 'Punit by condition' does not correlate with self-reported value similarity or any of the trust measures, except final trust in the AI condition. Mind perception is moderately correlated with value similarity and all trust measures, except WOA in the human condition. There are small correlations between decision time and all trust measures in the AI condition, but none in the human condition.

AI **Human**	Value similarity	Punitiveness	WOA	Final trust	Mind perception	Case trust	Punit by condition	Difference	Decision time
Value similarity	1	0.19*	0.41**	0.53**	0.34**	0.45**	-0.16†	-0.29**	0.09
Punitiveness	-0.15	1	-0.03	0.16†	0.44**	0.01	-0.16†	0.07	0.01
WOA	0.48**	-0.24†	1	0.50**	0.23*	0.67**	-0.11	-0.18†	0.22*
Final trust	0.70**	-0.09	0.51**	1	0.42**	0.75**	-0.24*	-0.06	0.23*
Mind perception	0.42**	-0.02	0.14	0.55**	1	0.39**	-0.09	-0.07	0.08
Case trust	0.64**	-0.06	0.64**	0.79**	0.34*	1	-0.14	-0.17†	0.20*
Punit by condition	0.01	-0.30*	0.13	-0.03	0.00	0.00	1	-0.11	0.00
Difference	-0.53**	0.09	-0.39**	-0.63**	-0.25†	-0.65**	-0.01	1	-0.17
Decision time	0.08	0.19	0.10	-0.09	0.08	-0.13	-0.18	-0.06	1

Note. † p < 0.1, * p < 0.05, ** p < 0.01.

Table 4. Pearson correlations of all measured variables in the AI condition (upper diagonal) and the human condition (lower diagonal).

Factor analysis

To determine the validity of the measured variables a confirmatory factor analysis (CFA) is conducted. Initially, all variables are included (value similarity, punitiveness, trust (11-item scale), and mind perception). The Kaiser-Meyer-Olkin Measure of Sampling Adequacy is 0.914. Bartlett's Test of Sphericity is significant ($\chi^2(465) = 4587.58$, $p < .001$). This indicates the sample is suitable for factor analysis. Analysis of the eigenvalues of the factors shows four factors with eigenvalues above one. Inspection of the corresponding Scree plot provided further evidence for a four-factor model. CFA was conducted on the four-factor model. Results show that all individual items significantly load onto their respective constructs, with a minimum standardized factor loading of 0.669. Fit statistics indicate the four-factor model fit is tolerable (RMSEA = 0.072, CFI = 0.916, TLI = 0.909, $\chi^2(428) = 803$) (Hu & Bentler, 1999). Inspection of the rotated factor loadings shows each item loads onto its respective scale, with minimal cross-loadings to the other scales (maximum cross-loading < 0.35). Factor loadings of all items can be found in the appendix.

Theory suggests that the 11-item trust scale consists of 3 distinct sub-scales: reliability, technical competence, and faith (Madsen & Gregor, 2000). However, inspection of the eigenvalues and Scree plot for these 11 items provides strong evidence for a single-factor structure. CFA is conducted including only the 11-item trust scale. A one-factor and a three-factor model are examined. For the one-factor model, all factor loadings are significant with a minimum value of 0.840. Model fit statistics are mixed (CFI = 0.950, TLI = 0.937, RMSEA = 0.107, $\chi^2(44) = 129$). CFI and TLI indicate a good fit, however, the RMSEA is too large. The fit of the three-factor model is better (CFI = 0.978, TLI = 0.971, RMSEA = 0.073, $\chi^2(41) = 78$). For the three-factor model, all individual items have significant loadings onto their respective sub-scales (minimum item loading = 0.871). However, further inspection of the factor loadings showed strong cross-loadings of most items onto multiple of the sub-scales. A two-factor model was additionally explored, however, this showed no improvement. The three theorized subcomponents are not distinctly present in the data. Therefore, the 11-item trust scale will be interpreted as a single-factor scale measuring trust. This scale does have good internal reliability (Cronbach's alpha = 0.96).

Influence of value similarity on psychological trust and WOA

The influence of perceived value similarity on psychological trust is analyzed through multiple linear regression. The DV is psychological trust measured at the end of all cases. The IVs are value similarity, advisor condition (human v AI), and advice condition (lenient v strict). Additionally, participant age, gender, and education were included as covariates. Variables were added in a stepwise manner, resulting in three separate models. The regression results are shown in Table 5.

	Model 1	**Model 2**	**Model 3**
Model variables	Coefficient (SE)	Coefficient (SE)	Coefficient (SE)
Intercept	18.8 (1.40)**	4.86 (2.28)*	5.35 (3.69)
Strict advice	0.15 (0.08)*	0.10 (0.06)	0.10 (0.06)
Human advisor	0.16 (0.08)*	-0.17 (0.18)	-0.15 (0.18)
Value similarity		0.51 (0.07)**	0.51 (0.07)**
Human advisor*Value similarity		0.31 (0.19)	0.27 (0.19)
Male			-0.05 (0.07)
Age			0.06 (0.07)
Education			-0.06 (0.06)
Multiple R^2	0.0473	0.384	0.405
F-statistic	F(166,2) = 4.12*	F(164,4) = 25.6**	F(159,7) = 15.48**
Observations	169	169	167
Note. † p < 0.1, * p < 0.05, ** p < 0.01. Coefficients are standardized.			

Table 5. Results of the linear regressions with psychological trust as the dependent variable.

Visual inspection of the residuals indicated a roughly normal distribution. Approximate normality was verified using the Shapiro-Wilk test (W = 0.99, p = 0.30). No significant outliers were present in the data. There were, however, potential problems of non-linearity and heteroscedasticity. Both these problems were resolved by applying a square transformation to the DV. The Ramsey RESET test indicated linearity in the transformed DV (F(2,157) = 1.61, p = 0.20). The Breusch-Pagan test supported that variance was homoscedastic after the transformation (p = 0.67). There was no indication of strong multicollinearity between any of the IVs or covariates (max VIF < 1.5).

Before adding any variables, the intercept-only model was inspected, which was significant (M = 21.9, SE = 0.84, p < 0.01). Model 1 contains only the experimental conditions. The advice and advisor condition significantly influence psychological trust, with the human advisor and strict advice being trusted more. However, the explained variance is low (R^2 = 0.0473). In model 2, value similarity and its interaction with the advisor condition are added. Value similarity positively influences trust. The addition of value similarity to the model shows a considerable increase in R-squared (R^2 = 0.384). In model 2 both experimental conditions are no longer significant. Additionally, there is a noticeable change in their regression coefficients, with the coefficient for advisor type changing sign. The addition of the demographic covariates in model 3 does not result in any notable changes.

The final model is significant with an R-squared of 0.405, a marginal increase compared to model 2. Value similarity has a significant positive effect on psychological trust in AI advice. This result supports H1(i). The non-transformed relation between psychological trust and value similarity is visualized in Figure 6 (left). Advisor type, the interaction between advisor and value similarity, and all of the included covariates were non-significant. Results show no evidence that psychological trust in an AI advisor is smaller than in a human advisor. Thus we find no support for H3(i). Similarly, there is no significant evidence that the influence of value similarity on psychological trust is stronger for a human advisor than for an AI advisor, thus, H2(i) is not supported.

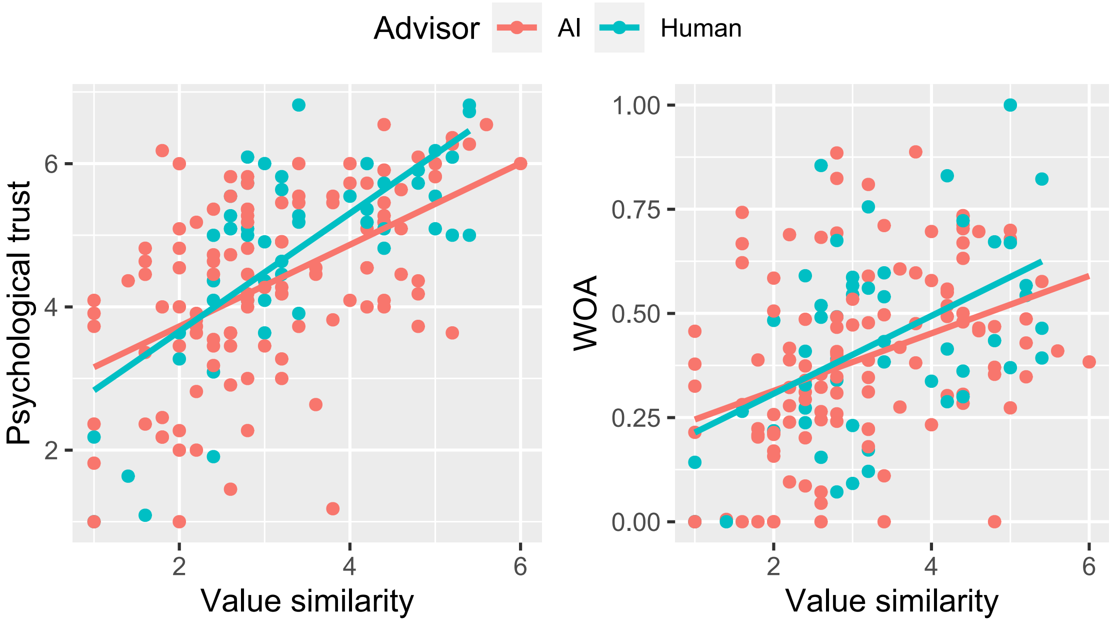


Figure 6. (left) Relationship between psychological trust and perceived value similarity. (right) Relationship between WOA and perceived value similarity. Shown for both the AI and human advisor conditions.

To test whether the positive influence of value similarity on trust results in a behavioral change in the form of compliance, another regression with the same IVs and covariates was performed with WOA as the DV. The regression results are presented in Table 6.

	Model 1	**Model 2**	**Model 3**
Model variables	Coefficient (SE)	Coefficient (SE)	Coefficient (SE)
Intercept	0.367 (0.03)**	0.155 (0.05) **	0.174 (0.08)*
Strict advice	0.22 (0.08)**	0.18 (0.07)*	0.18 (0.07)*
Human advisor	0.09 (0.08)	-0.10 (0.21)	-0.06 (0.20)
Value similarity		0.38 (0.08)**	0.43 (0.08)**
Human advisor*Value similarity		0.18 (0.21)	0.12 (0.21)
Male			-0.06 (0.07)
Age			-0.09 (0.07)
Education			0.05 (0.07)
Multiple R^2	0.054	0.224	0.265
F-statistic	F(166,2) = 4.77**	F(164,4) = 11.84**	F(159,7) = 8.19**
Observations	169	169	167
Note. † p < 0.1, * p < 0.05, ** p < 0.01. Coefficients are standardized.			

Table 6. Results of the linear regressions with WOA as the dependent variable.

To meet the regression assumptions, the transformation x -> $x^{0.9}$ was applied to the DV. Inspection of the WOA distribution identified 11 participants with an average WOA of 0. This means that they did not, for any of the 14 cases, change their initial decision. This could indicate a lack of engagement with the task. The regression was run with and without these cases included. No notable difference in outcomes resulted from this, therefore the participants were included in the final analysis.

The intercept-only model of WOA was significant (M = 0.408, SE = 0.017, p < 0.01). In model 1, WOA is greater for strict advice than for lenient advice. The addition of value similarity and its interactions in model 2 shows a positive influence of value similarity on WOA. Contrary to the analysis of psychological trust, advice type remains a significant predictor when value similarity is included. The inclusion of the covariates in model 3 does not result in any meaningful changes to the model.

Overall, the results show that value similarity positively affects WOA. The relation between value similarity and WOA is visualized in Figure 6 (right). As expected, value similarity increases compliance with advice, supporting H1(ii). Similarly to the results for psychological trust, there is no significant difference in WOA between human and AI advisors. Thus no support for H3(ii) is found. The interaction between value similarity and advisor was also non-significant, indicating that the influence of value similarity on compliance is not stronger for a human advisor. Therefore, H2(ii) is also not supported. Contrary to psychological trust, WOA is positively influenced by advice type, with strict advice being complied with to a greater extent.

Value similarity based on punitiveness

The previous analysis of the relationship between value similarity and trust is based on self-reported value similarity measured at the end of all 14 tasks. Another method of analyzing value similarity is through a comparison of participants' punitive attitudes to the punitiveness of the advisor. As per the initial data analysis proposal, the continuous punitiveness variable is split into two categories (high and low). A dichotomous value similarity variable is created by pairing high or low punitiveness with the strict or lenient advisor condition. The effect of value similarity is tested through two 2x2 ANCOVAs with psychological trust and WOA as DVs and value similarity and advisor as IVs. Age, gender, political affiliation, and education are included as covariates.

Results show a significant effect of value similarity on psychological trust (F(1,160) = 5.75, p = 0.02, η2 = 0.03). Additionally, age is a significant predictor of psychological trust (F(1,160) = 6,60, p = 0.01, η2 = 0.04). The test showed no significant influence of the advisor being human, nor any of the other covariates, on psychological trust. Value similarity, advisor, nor any of the covariates had a significant influence on WOA.

This method of splitting the continuous variable into two categories was used in previous similar studies and thus allows easier comparisons to previous literature. However, splitting the variable in this way means a considerable amount of information is lost. Therefore, the influence of the interaction between participant punitiveness and advice condition on psychological trust and WOA is additionally analyzed through multiple linear regression. As before, a square transformation is applied to psychological trust to adhere to the regression assumptions. The regression models with psychological trust as the DV are presented in Table 7.

	Model 1	**Model 2**	**Model 3**
Model variables	Coefficient (SE)	Coefficient (SE)	Coefficient (SE)
Intercept	18.8 (1.40)**	18.6 (4.80)**	20.1 (5.68)**
Strict advice	0.15 (0.08)*	-0.20 (0.26)	-0.29 (0.27)
Human advisor	0.16 (0.08)*	0.30 (0.11)**	0.29 (0.11)**
Punitiveness		-0.02 (0.13)	-0.11 (0.13)
Punitiveness*Strict		0.50 (0.29)†	0.60 (0.29)*
Punitiveness*Strict*Human		-0.25 (0.12)*	-0.28 (0.12)*
Education			-0.08 (0.08)
Age			0.19 (0.08)*
Male			-0.16 (0.08)*
Multiple R^2	0.0473	0.091	0.138
F-statistic	F(166,2) = 4.12*	F(163,5) = 3.25**	F(158,8) = 3.18**
Observations	169	169	167
Note. † p < 0.1, * p < 0.05, ** p < 0.01. Coefficients are standardized.			

Table 7. Results of the regression models analyzing the interaction between punitiveness and advice condition on psychological trust.

Note that the intercept-only models and model 1 for both psychological trust and WOA are the same as in the previous regression analysis, with self-reported value similarity as the IV. Therefore, they will not be discussed again. In model 2, punitiveness and its interactions with the advice and advisor condition are added. Notably, the positive influence of the human advisor remains when the interactions are added.

In model 3 there is a significant interaction effect between participant punitiveness and advisor condition on psychological trust. This shows that people with stronger punitive attitudes report greater trust in a strict advisor. This effect is visualized in figure Figure 7 (left). Additionally, the three-way interaction between punitiveness, advice condition, and advisor condition is also significant. This shows that the interaction between punitiveness and the advice condition is weaker for a human advisor. Furthermore, the covariates age and gender are significant. Older participants have more trust in their advisors and male participants have less trust. Overall, these results are similar to the results of the ANCOVA. The final regression model is significant with an R-squared of 0.138. This is considerably lower than the R-squared of 0.405 found for the model where self-reported value similarity was used as the main predictor of trust. This suggests that the self-reported value similarity measure is a better predictor of psychological trust.

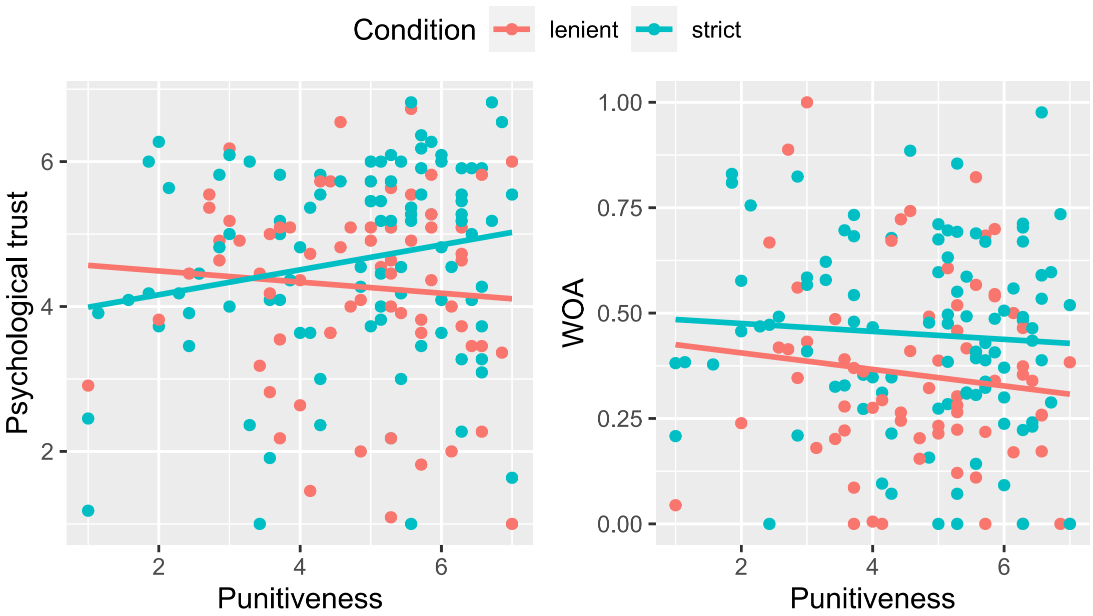


Figure 7. (left) Condition-dependent relation between psychological trust and punitiveness. (right) Condition-dependent relation between WOA and punitiveness.

Overall, these results support H1(i). Value similarity, operationalized as the interaction between punitiveness and advice type, positively affects psychological trust in an AI advisor. Trust in the AI advisor is significantly lower than trust in the human advisor, supporting H3(i). Contrary to expectation, the three-way interaction implies that the influence of shared values is weaker for a human advisor than for an AI. This result is the opposite of H2(i).

Again, the same regression models were also constructed with WOA as the DV. The resulting models are shown in Table 8.

	Model 1	**Model 2**	**Model 3**
Model variables	Coefficient (SE)	Coefficient (SE)	Coefficient (SE)
Intercept	0.33 (0.03)**	0.38 (0.10)**	0.40 (0.12)**
Strict advice	0.22 (0.08)**	0.13 (0.26)	0.04 (0.27)
Human advisor	0.09 (0.08)	0.30 (0.11)**	0.29 (0.11)**
Punitiveness		-0.11 (0.13)	-0.13 (0.13)
Punitiveness*Strict		0.23 (0.29)	0.33 (0.30)
Punitiveness*Strict*Human		-0.31 (0.12)*	-0.34 (0.12)**
Education			0.01 (0.08)
Age			0.03 (0.08)
Male			-0.13 (0.08)
Multiple R^2	0.054	0.099	0.120
F-statistic	F(166,2) = 4.77**	F(163,5) = 3.57**	F(158,8) = 2.68**
Observations	169	169	167
Note. † p < 0.1, * p < 0.05, ** p < 0.01. Coefficients are standardized.			

Table 8. Results of the regression models analyzing the interaction between punitiveness and advice condition on WOA.

Note that, contrary to the previous analysis, no transformation was applied to WOA. Overall, the results of the three models for WOA are similar to the ones with psychological trust as the DV. The main difference is that the punitiveness and advice condition interaction is non-significant. Additionally, none of the covariates has a significant influence on WOA. The final model is significant with an R-squared of 0.120. Again, this is considerably lower than the R-squared of the model using self-reported value similarity as the predictor (R-squared = 0.265).

These results do not support H1(ii). Alignment between participants' punitive attitudes and the advice they received did not significantly increase WOA. WOA was greater for advice provided by a human than an AI, supporting H3(ii). However, similar to the psychological trust result, the significant three-way interaction indicates that the influence of shared beliefs is stronger in the AI condition, opposing H2(ii).

Overall, looking at the results of the regression for both psychological trust and WOA, there are clear discrepancies between the predictions based on the constructed variable of value similarity and the self-reported measure. The variance explained by the models relying on the constructed variable is much lower. Additionally, when using the constructed variable, the results differ between WOA and psychological trust. This is contrary to the strong correlation that was found between WOA and self-reported trust measures. Given that the constructed and the self-reported variable do not significantly correlate with each other either, it is questionable whether the constructed variable is an adequate measure of value similarity.

Trust development over time

Trust development is analyzed using mixed effects regression with trust as the DV and trial number, error, sentence difference, advisor, and value similarity as IVs. Trust is based on the three-item scale administered after every individual case. The variable trial number refers to when the case was presented to the participant, ranging from 1 – first trial to 14 – last trial. The variable 'error' is the relative difference between the advice and the judge's sentence. The error ranges from 0% (advice is the same as the judge's sentence) to 25% (advice is 25% higher or lower than the judge's sentence). The variable 'difference' is the absolute difference in months between a participant's initial sentence and the advice provided to them. Variables are added in a stepwise manner. The human advisor is excluded from the analysis as the interest is only in trust development in the case of AI advice. The regression results of the three examined models are shown in Table 9.

	Model 1	**Model 2**	**Model 3**
Model variables	Coefficient (SE)	Coefficient (SE)	Coefficient (SE)
Intercept	19.9 (1.76)**	9.72 (2.95)**	12.1 (5.25)*
Strict advice	0.13 (0.08)†	0.07 (0.07)	0.06 (0.07)
Value similarity		0.36 (0.07)**	0.36 (0.08)**
Trial number		0.06 (0.01)**	-0.01 (0.04)
Difference		-0.06 (0.02)**	-0.17 (0.05)**
Error		-0.15 (0.01)**	-0.03 (0.04)
Trial number*Value similarity			0.09 (0.05) †
Difference*Value similarity			0.11 (0.05)*
Error*Value similarity			-0.15 (0.04)**
Age			0.05 (0.07)
Male			-0.19 (0.07)**
Education			-0.03 (0.07)
Marginal R squared	0.018	0.179	0.237
Observations	1649	1649	1621
Number of groups	118	118	116
Note. † p < 0.1, * p < 0.05, ** p < 0.01. Coefficients are standardized.			

Table 9. Mixed effects regression results with psychological trust as the target variable. Only the AI condition data is included in the analysis.

As in the previous two analyses, a square transformation is applied to the DV. Within the difference variable, there were 3 outliers with significant leverage on the results. These participants gave prison sentences of over a thousand months. It could be the case that they were trying to give 'life sentences' this way. However, based on the rest of their answers, which were not significantly more strict than other participants, it is assumed that these answers are typing mistakes. The three data points have been excluded from the analysis. Inspection of the error variable showed that the maximum error of the advice was 25%. The maximum error in the experiment was supposed to be fixed at 20%. This difference is due to the advice for one of the lower prison sentences being rounded up by one month in the randomization calculations. Given that these are the actual values participants saw during the task, they are included in the analyses.

The intercept-only model of trust is significant (4.36, SE = 0.14, $p < 0.01$). Model 1, including only the advice condition, shows no significant difference between a strict and lenient AI. In model 2, trial number and the other measured variables are added, all of which are significant predictors of trust. Similarly to previous results, value similarity significantly positively influences trust. Contrary to expectations, trial number has a positive influence on trust as well. However, after the inclusion of the value similarity interactions and demographic variables in model 3, the effect of trial number disappears. Further inspection of trial number compared to trust shows that, while statistically significant in model 2, the total change in trust between trials 1 and 14 is only +0.16 (trust is measured on a 7-point scale), which is negligible. It was also tested if the effect of trial number differed between conditions. However, in both the lenient and strict condition trial number and its interactions are non-significant in model 3.

In model 2, error negatively influenced trust. When advice deviated further from the sentence handed out by the judge, trust in the advice decreased. This suggests that participants partly based their trust evaluation on a comparison between their answer and the sentence of the judge. After inclusion of the value similarity interactions in model 3, the effect of error is reduced and no longer significant. There is a significant interaction between error and value similarity. This is shown in Figure 8 (left). When value similarity with the advisor is greater, the negative effect of error on trust is stronger. This shows that participants were less trusting of advice that deviated from the decision of the judge when they perceived the advisor as having similar values as them.

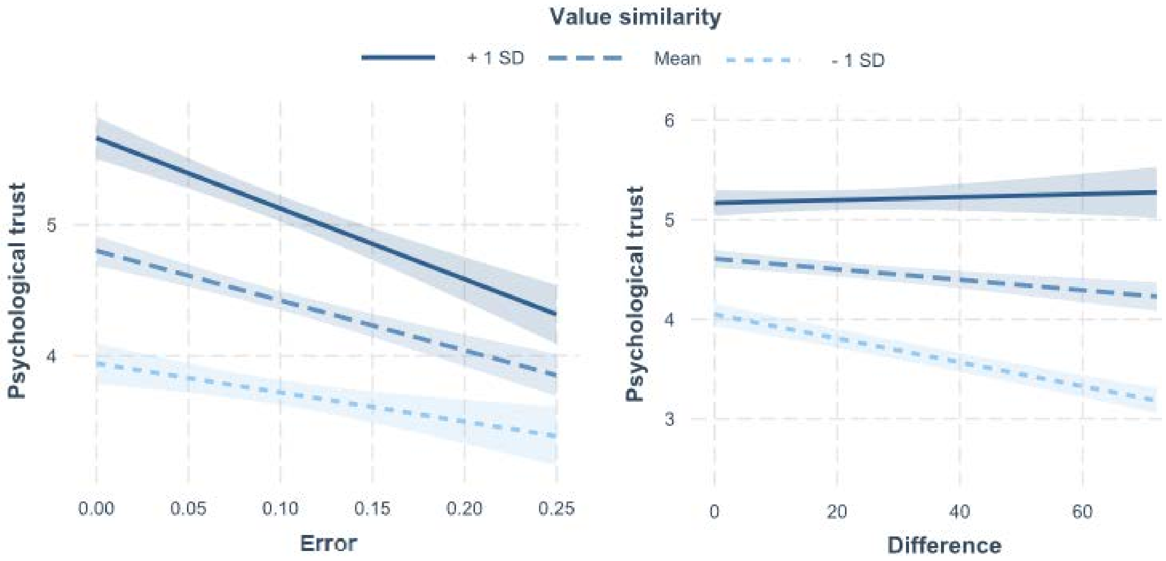


Figure 8. (left) The relation between error and psychological trust for varying levels of value similarity. (right) The relation between sentence difference and psychological trust for varying levels of value similarity.

The difference between a participant's initial sentence and the advice provided to them negatively influenced trust. This implies that participants' judgments are partly formed by comparing the received advice to their own judgment. Similarly to the influence of error, the influence of sentence difference on trust in advice is moderated by value similarity. This interaction is visualized in Figure 8 (right). The negative impact of the difference between a participant's sentence and the advice is smaller when value similarity between the participant and the advisor is greater. This result shows that people are more trusting of advice that differs from their initial decision when they perceive the advisor as having similar values.

Overall, the results of the final model support H1(i), value similarity positively affects psychological trust. This result adds to the results of the previous two analyses by showing that value similarity increases trust in advice on a case-by-case basis. It also provides further evidence that value perceptions played a role in trust formation during the task, given that value similarity was measured after the tasks. It was expected that value similarity would moderate the effect of trial number on trust. However, the results show no significant interaction effect. Therefore, H4(i) is not supported. From the included covariates only gender was significant. Male participants reported lower trust in the AI on a case-by-case basis.

Another mixed effects regression was performed with the same predictors and interactions but with WOA as the dependent variable. Error and the interaction between value similarity and error are excluded from this analysis. This is done because the participants were not aware of the judge's verdict at the time they made their final decision, thus they are unaware of the error. The results of the regression models are shown in Table 10.

	Model 1	Model 2	Model 3
Model variables	Coefficient (SE)	Coefficient (SE)	Coefficient (SE)
Intercept	0.31 (0.03)**	0.02 (0.06)	0.24 (0.11)*
Strict advice	0.20 (0.05)**	0.16 (0.05)**	0.15 (0.05)**
Value similarity		0.24 (0.05)**	0.15 (0.07)*
Trial number		0.06 (0.02)**	0.01 (0.06)
Difference		0.08 (0.02)**	-0.14 (0.08)†
Trial number*Value similarity			0.07 (0.07)
Difference*Value similarity			0.23 (0.07)**
Age			-0.06 (0.05)
Male			-0.11 (0.05)*
Education			-0.03 (0.05)
Marginal R squared	0.039	0.094	0.095
Observations	1481	1481	1454
Number of groups	118	118	116
Note. † p < 0.1, * p < 0.05, ** p < 0.01. Coefficients are standardized.			

Table 10. Mixed effects regression results with WOA as the target variable. Only the AI condition data is included in the analysis.

WOA values outside the 0 – 1 range have been excluded from the analysis. In total 171 observations (out of 1652) were excluded. 54 of these were due to participants giving the same initial sentence as the AI advised. Possible explanations for the rest of the excluded values are discussed later.

The intercept-only model of WOA is significant (0.39, SE = 0.02, p < 0.01). In model 1, there is a significant positive influence of strict advice on WOA. This effect remains after the inclusion of the measured variables in model 2 and the interactions and demographics in model 3. Again, while trial number significantly influences WOA in model 2, the effect is negligible and disappears completely in model 3. Notably, in model 2, there is a significant positive effect of sentence difference on WOA. This implies that participants complied to a greater extent with advice that differed more from their initial decision. This is opposite to the finding for psychological trust. The reason for this finding is likely due to the nature of the WOA measure. As it measures relative adjustments, when differences are larger, greater WOA values are to be expected. The inclusion of the value similarity interaction in model 3 results in the effect of difference becoming negative. This is the expected effect direction. However, the effect is no longer significant.

Overall, the WOA results follow a similar pattern as those for psychological trust. The influence of value similarity on WOA is positive. Participants were more likely to comply with advice from an AI that they perceived as having similar values to theirs. This supports H1(ii). Contrary to expectation, neither trial number nor the interaction between trial number and value similarity influences WOA. No evidence is found that trust develops over time, nor that this development is affected by value similarity. Thus, H4(ii) is not supported. Of the included covariates, only gender is significant. WOA was lower for male participants.

Decision-making time as a measure of trust

Decision-making time is examined as another behavioral measure of trust in addition to WOA. Decision time is defined as the time in seconds that a participant took to make their final decision after receiving advice. A mixed effects regression was performed with the same predictors and interactions as for the WOA analysis. Error and the interaction between value similarity and error are again excluded as participants were not aware of the judge's verdict when making their decision. The results of the regression models are shown in Table 11.

	Model 1	**Model 2**	**Model 3**
Model variables	Coefficient (SE)	Coefficient (SE)	Coefficient (SE)
Intercept	8.01 (0.48)**	9.63 (0.96)**	6.8 (1.59)**
Strict advice	0.08 (0.07)	0.06 (0.07)	0.07 (0.06)
Value similarity		0.07 (0.07)	-0.12 (0.08)
Trial number		-0.32 (0.02)**	-0.31 (0.05)**
Difference		0.13 (0.02)**	-0.30 (0.06)**
Trial number*Value similarity			-0.00 (0.06)
Difference*Value similarity			0.45 (0.06)**
Age			0.39 (0.07)**
Male			-0.04 (0.07)
Education			-0.02 (0.07)
Marginal R squared	0.01	0.12	0.26
Observations	1562	1560	1533
Number of groups	118	118	116
Note. † p < 0.1, * p < 0.05, ** p < 0.01. Coefficients are standardized.			

Table 11. Mixed effects regression results with decision time as the target variable. Only the AI condition data is included in the analysis.

Inspection of the decision time variable shows that there are many extreme outliers in the data (Mean = 11, Median = 7, Range 2 - 292). This is unsurprising for an online survey without time constraints. Participants were likely distracted for a couple of minutes after which they continued the study. There is no way to accurately determine all decision times that are partly influenced by real-world distractions. Therefore, a cutoff point is set at 25 seconds. This is roughly three times the average value after exclusion of outliers. Given the task, this is deemed a fair amount of time for completion of the question assuming no distractions.

There is a significant negative effect of trial number on decision time. This can partly be explained by participants becoming more familiar with the task and therefore giving faster answers. However, this is also a possible indicator that participants were getting bored and trying to get through the tasks quicker toward the end. The significant positive influence of age on decision time shows that older participants took longer to complete the tasks.

In model 2, difference positively influences decision time. This indicates that people took longer to decide when the AI's advice differed more from their initial decision. There is no significant effect of value similarity on decision time. Inclusion of the interaction effects in model 3 does show that there is a significant positive interaction between value similarity and difference. Notably, upon inclusion of this interaction, the main effect of difference changes sign, becoming negative. A possible interpretation of these opposite effects is that people are more willing to consider advice that differs from their initial decision when the advice aligns with their own beliefs. Overall, larger differences between advice and initial decision result in people taking less time to evaluate the advice, and they are even quicker to dismiss the advice when it does not match their values.

Perception of values

As an additional exploratory part of this research mind perception was measured. As expected, mind perception is significantly greater for the human advisor (M = 5.30, SD = 0.81) than the AI (M = 2.76, SD = 1.21) (F(1,166) = 187, p < 0.001, η2 = 0.53). A regression with mind perception as the DV and all demographic variables as IVs found that mind perception of the AI was lower for higher educated (β = -0.18, SE = 0.09, p = 0.047) and male participants (β = -0.53, SE = 0.23, p = 0.022).

Shank et al. (2021) found that the difference in value attributions between humans and AI is mediated through mind perception. As value attribution is a necessity for value similarity to exist, it is expected that a similar mediation is present in the case of value similarity. It is reasoned that if mind perception plays a role in value similarity with the AI, that this is evidence that value similarity with the AI is not solely the result of viewing the AI's output as an expression of the values of its developers.

A mediation analysis with 10,000 bootstraps was conducted to test this. Results are shown in Figure 9. In the mediation model, the AI advisor is taken as the baseline. The indirect effect of advisor type on value similarity through mind perception is significant (0.962, 95%CI = [0.515, 1.43], p < 0.001). The direct effect of advisor type on value similarity is also significant (-0.739, 95%CI = [-1.29, -0.17], p = 0.01), but it has the opposite sign. The sign of the direct and indirect being opposite partly explains why the total effect of advisor type is insignificant (0.223, 95%CI = [-0.16, 0.61], p = 0.24). Looking at only the AI condition, the correlation between mind perception and value similarity remains significant (0.34, p < 0.001). The overall result of the mediation is in line with prior literature stating that mind perception is an important aspect of perceived value similarity. The significant relation between mind perception and value similarity with the AI supports the idea that participants perceived the AI itself as a value-based actor.

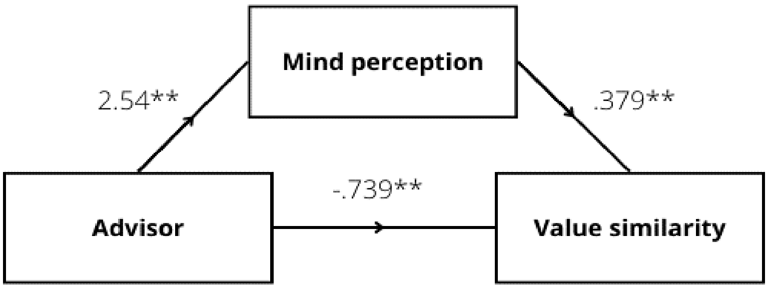


Figure 9. Results of the mediation analysis with the advisor (AI = 0, human = 1) as IV, value similarity as DV, and mind perception as the mediator.

Qualitative results

In addition to the quantitative measures, two open questions were included. Participants were asked to further elaborate on their answers as to why they felt the advisor was lenient or strict and if they thought the task was realistic. 137 participants answered one or both of the open questions.

The main difference in the interpretation of the punitiveness of the advisor is who a participant compares their advisor to. Either to themselves or the judge. Some participants explained that they chose to base their answer on a comparison to the judge as they themselves lack legal expertise. Others clearly state that they think the judge is wrong or politically biased and therefore compare the advice to their own sentences. These statements indicate a difference in the influence of deviation from the judge's sentences and deviation from the participant's answers on trust in the advisor. These answers support the quantitative findings. This discrepancy further shows that the responses to the manipulation check are potentially unreliable. A participant who was aware that their advisor was strict (compared to the judge) may still have answered that they thought the advisor was lenient (compared to them) and vice versa.

A few participants said they felt the AI made decisions purely based on build-in knowledge of prior verdicts, and that lenience or strictness was thus not applicable to the AI's output. This implies that these participants did not perceive the AI to make value-based decisions. One participant stated that the AI was missing the crucial element of human understanding, emotion, and empathy in its decision-making. Another participant said the AI was perhaps unable to understand the emotional harm caused to the victims. Other participants similarly stated that the AI was not considering the feelings of the defendant and was unable to understand a person's mental state. These comments are in line with the quantitative findings showing that both value similarity and mind perception are greater for a human advisor.

Some participants specifically mentioned the beliefs of the advisor and compared these to their own beliefs. Others instead explained the behavior of the AI purely based on the beliefs and ideas of its programmers. This interpretation is unaccounted for in our research method. No distinction is made between value similarity with the AI as an entity in and of itself and value similarity with the AI's developers.

Participants indicated their opinions vary greatly for different types of crimes (e.g., murder, drugs, or robbery), while the advisor's opinion did not seem to do so. Participants indicated that they felt the advisor should have made a distinction between violent and non-violent crimes. Concerning this, multiple participants indicated that additional rehabilitation measures should be considered instead of only prison sentences and giving defendants fines instead of jail time. Overall, there was a clear sentiment among participants that the advisor in both conditions was too lenient. The same sentiment was held for the sentences handed out by the judge. Many participants indicated that they were sorry and disappointed in the legal system, stating that especially perpetrators of violent crimes should be incarcerated for longer to protect the public. These sentiments are in line with the quantitative measure of punitiveness that indicated the participant sample was on average quite punitive (M = 4.79, 1-7 range). This overall feeling that the advisors were not strict enough may be explained by some participants stating that they purposefully gave double their desired sentence. These participants indicated that they know that people in the UK tend to serve about half their allotted time (the same is true for the Dutch sentences the cases are based on) and therefore gave higher sentences. One participant specifically indicated they were unsure if the advisor was considering early release. This behavior could damage the interpretability of the results as it makes lenient people (low scores on punitiveness) seem stricter in their answers.

Discussion

The main aim of this research was to analyze the influence of value similarity on psychological and behavioral trust in biased AI advice. Additionally, the difference in trust between an AI and human advisor, trust development over time, and mind perception of AI were investigated.

The influence of value similarity on trust and compliance

The results suggest that both overall trust in the AI advisor, as well as, trust on a case-by-case basis is positively affected by value similarity. This finding is in line with previous AI value similarity studies (Mehrotra et al., 2021; Yokoi & Nakayachi, 2019, 2021a, 2021b). The present research expands on these findings by showing that value similarity also affects trust in biased advice. The fact that deviation from the judge's sentence negatively influences trust shows that people did perceive the AI's offset as an error. Thus, the positive influence of value similarity implies that people's trust is greater when they agree with the bias of the AI. A potential explanation is that people are more forgiving of errors when those align with their beliefs. Value similarity was also found to positively influence compliance with advice. This is in line with previous studies that found value similarity to positively influence behavioral intention (Liu & Moore, 2022; Mehrotra et al., 2021). The present study further adds to this by showing that trust due to value similarity results in advice adoption.

While the self-reported measure of value similarity produced clear results, the more objective measure, namely the interaction between the advice condition and the punitive beliefs of the participants, was less clear. The objective measurement of value similarity positively influenced trust but did not affect compliance. This result is opposite to the findings of Liu & Moore (2022), who found political belief alignment, measured similarly, to influence behavioral intention but not trust positively. An explanation could be that the objective measure is not valid or reliable. Mehrotra et al. (2021) found that it is difficult to match participants' measured values with AI value profiles. The perception of value similarity with an AI is complex and based on more than a single value or belief. Future research is required to gain a better understanding of the value similarity perceptions of AI.

The impact of advice accuracy and difference on trust

Alongside value similarity, the other significant predictors of trust and compliance were advice accuracy (relative to the judge) and advice difference (relative to the participant). This implies that participants based their trust and compliance on a comparison of the AI's advice to the judge's sentence and their own decision. Lower trust in less accurate advice is in line with past findings (Lockey et al., 2021). It is interesting that the difference between a participant's decision and the AI's advice negatively impacted trust. Given the low legal expertise of the participants, it would be expected that receiving completely different advice is seen as an indication that their initial decision was unrealistic. Difference having a negative effect despite this indicates that prison sentencing was partly seen as a subjective task.

Value similarity interacted with accuracy and difference. The hypothesis that value similarity would moderate the negative effect of inaccurate advice is unsupported. Instead, value similarity seems to result in a higher base level of trust. Seeing an AI make mistakes still has a similar negative effect on trust, regardless of whether those errors align with personal beliefs.

The results show an interesting contrast in the interaction between difference and value similarity. Value similarity alleviates the negative effect of the difference between a person's own decisions and the advice they received. This result suggests that when people compare the AI's recommendation to their own decision instead of the verdict of the judge, value similarity is more influential. Previous research has shown that domain experts are less trusting of advice and are more likely to compare recommendations to their own knowledge (Logg, 2017). Thus, it might be the case that the effect of value similarity is stronger for people with greater task expertise. Unfortunately, due to the low average legal expertise of the sample, this idea could not be tested.

Differences between the human and AI advisor

As expected, results show that overall trust and compliance are more significant for a human than an AI advisor. However, the positive effect of human advice disappears when self-reported value similarity is included. Human advice is generally preferred when decisions pertain to a personal or ethical situation. Responses confirm that prison sentencing was viewed as an ethical decision. In a study comparing a biased human to a biased AI, Langer et al. (2022) found that participants had greater expectations of a human advisor than an AI. They argue that the positive effect of human advice is diminished due to people being more condemning of advice that clashes with their personal beliefs when it comes from a human as opposed to an AI. This might have been the case in the present study.

Another explanation is that the distinction between human and AI conditions was not pronounced enough. The only difference is changing the word 'expert' to 'AI'. Yokoi & Nakayachi (2021b) found a similar result when distinguishing between humans and AI only by name. Possibly the distinction between a human and AI is oversimplified in the present study, especially since the advice is provided digitally through an online survey, which may further take away the human component. However, the mind perception measure does show that people clearly understood that they were receiving advice from a non-human.

The study results do not support the hypothesized interaction between value similarity and advisor type. For self-reported value similarity, the effect of value similarity was not significantly different between humans and AI. For the constructed measure of value similarity, the interaction was opposite to expectations; value similarity played a minor role in trust in the human advisor. However, as discussed before, the accuracy of this variable as a means of measuring value similarity is questionable. These findings are again similar to those of Yokoi & Nakayachi (2021b).

Trust development over time

It was expected that trust in the AI's advice would decrease over time due to participants witnessing the imperfections of the AI. However, we found little evidence that repeated interactions with the AI impact psychological or behavioral trust. There was no meaningful difference between initial and later trust. Furthermore, no significant trust development was observed after inclusion of demographic variables. A possible explanation for this is the constant behavior of the AI. The current study does not include a single salient error in an overall accurate system, such as in the work of Renier et al. (2021). Instead, the AI has a constant small but noticeable bias. It was expected that seeing the AI occasionally make a considerable deviation from the judge's verdict (e.g., an extra year) would be sufficient to impact participants' trust. While this was found to be true on a case-by-case basis, there was no meaningful change in average trust levels throughout the experiment. However, the results of Dietvorst et al. (2015) show that trust can be negatively impacted even when there is not a single obvious error. Another reason why no trust development is present may be the high advice accuracy compared to the participants' own decisions. Participants may have noticed the occasional larger deviations of the AI, but still chose to rely upon this advice, given it was more accurate than their own decisions.

It was hypothesized that the similarity between the AI's bias and the participant's values would moderate the development of trust over time. It was expected that holding beliefs that correspond to that AI's bias would result in the inaccuracies being more tolerable, and the opposite when bias and beliefs misalign. Again, the results do not support this expectation. It may be that due to the relatively short duration of the task participants did not yet form a clear trusting attitude toward the AI. Instead, they might have evaluated trust on a case-by-case basis. The significant interactions between value similarity and case-wise errors and differences in advice support this idea.

Perception of values in AI advice

As an additional part of this research, the influence of mind perception on value similarity was examined. The obvious expectation that mind perception is greater for humans than for AI was met. However, people still perceived AI as having an agentic mind to some degree. Moreover, mind perception of the AI was positively related to value similarity. This result is in line with Cvetkovich's (2013) claims that mind perception is a prerequisite for value similarity. Furthermore, it expands on the finding by Shank et al. (2021) that mind perception influences value attribution to sophisticated AI by showing that this attribution is still present when the system is presented more straightforwardly (it just gave a number through an online survey in the current study). Overall, these results somewhat alleviate the concerns that values would not be perceived in the research setup.

Limitations

The trust definition proposed by Lee & See (2004) states that vulnerability is a crucial aspect of trust. A mistake should have some negative consequences for a participant to be vulnerable. In essence, prescribing prison sentences to criminals meets this requirement had the participants' sentences actually been applied. Since this is not the case, it is unlikely that participants felt their decisions were consequential. It was asked of participants to try and immersive themselves in the task and treat it as if they were handing out the sentences. However, whether this resulted in participants experiencing some vulnerability is unknown.

The experiment relied on a decision-making task with no upper bound. That is to say; there was no upper limit on how high a sentence could be. This lack of an upper limit led to multiple extreme outliers in the data. While some of these outliers were probably typing mistakes, participants may have intended these high sentences to be life sentences. Letting participants give their answers in a range (e.g., 0 – 100 months) would be better. A range prevents problems with the interpretation of outliers and provides participants with less legal knowledge a frame of reference for their decisions.

Some participants indicated they were adding on additional months to compensate for probation. Suppose a follow-up study uses a similar setup. In that case, it is important to inform participants that the full sentence will be served, and that probation should be neglected in the sentencing decision. Additionally, it would be beneficial to explain that in none of the cases, any additional punishments (e.g., fines or community service) were a (major) part of the sentence.

The responses made it apparent that many participants were handing out sentences in full years and then converting this to months (12, 24, 36, etc.). Calculating the total months of a specific number of years and months combined can be time-consuming and annoying. However, given that most Dutch sentences are a couple of years, rounding to full years makes answers much less accurate. The participants might have been able to give better judgments if they could express their answers more intuitively in years and months. Therefore, it is recommended that if a similar task were to be used in the future, answers could be given in years and months (e.g., four years and nine months).

We measured punitive attitudes once after the trials, but punitive attitudes may have varied between case descriptions. Adriaenssen & Aertsen (2015) have highlighted how punitive attitudes can vary significantly based on the type of crime and the personal characteristics of the victim and perpetrator. Answers to the open questions confirm that many participants agreed with this. Ideally, a separate measure of punitiveness is used for each crime. In the present study, this was not possible due to time constraints. Additionally, it is likely that having participants repeatedly answer the same questionnaire leads to fatigue and reduces the quality of the responses. To still try and combat these concerns, we chose a general scale of punitiveness with no focus on a specific type of crime or group and ensured a broad range of cases. Nevertheless, the type of cases and people's opinions on them may not have aligned with the general punitiveness measure resulting in it being an unreliable measure of value similarity on a case-by-case basis. This could explain why we find no significant correlation between self-reported value similarity and the constructed measure based on punitiveness.

Additionally, as the punitiveness measure was administered after the experiment, participants' answers could have been affected by reading the crime descriptions and seeing the sentences. This possibility may explain why the politically left-leaning participants in our sample were more punitive than expected. However, positioning the punitiveness measure at the end of the task may still be preferred. Inquiring about a person's punitiveness and then letting them work with a biased advisor while continuously asking about their trust may alert them to the experiment's goal.

The present study made use of a UK sample. The crime descriptions were based on crimes committed in The Netherlands, and a Dutch court made the sentencing decision based on Dutch law. We argue that there is considerable overlap between the Dutch and UK legal systems and that the jail sentences handed out are comparable in duration. Furthermore, given the sample's low average level of legal expertise, it is unlikely that participants would have based their estimates on the intricacies of their national legal system. However, participants may still have had underlying assumptions about UK prison sentencing that are at odds with the Dutch legal system. Ideally, a sample of only Dutch participants would be used.

Participants were tasked with determining a prison sentence and then shown the sentence of the judge. A participant in the pre-test (before the pilot) indicated that they were trying to get their final answer to be as close to the judge as possible. This behavior would mean that value similarity would be inconsequential to their decision-making process. Their decision would be based solely on how close the advice was to the 'correct' answer instead of whether they agreed with the advice. To prevent participants from trying to get close to the judge’s sentence, they were instructed to view the judge's decision as a reference to get a better idea of a common sentence in the legal system. Participants were also asked to try and envision that the prison sentences they were handing out would be the true prison sentence that the defendant would be subjected to. At the end of the experiment, participants were asked if they were able to immersive themselves in the task and if they could envision their sentences being handed out. On average, they indicated they could (M = 6.5, range 1-10). Responses to the open questions further showed that people did not just base their decision on the judge's sentence but took the advice and their beliefs into account. Furthermore, no participant indicated they were trying to get their answer as close to the judge as possible. However, they were also not specifically probed on this. Ideally, the task would not involve a 'correct' answer. We chose a task that did include this, in this experiment the judge's verdict, because it makes the task doable for non-experts. Having lay participants try to answer legal questions with no frame of reference would likely make them completely reliant on the advice, irrespective of their personal beliefs.

WOA is a commonly used measure of compliance in judge-advisory tasks. However, the measure has limitations, which also became apparent in the present research. There were multiple instances of WOA falling outside the 0-1 range. This means that a participant changed their answer in the opposite direction of the advice. This could be indicative of a participant realizing the system is biased and trying to get as close to the judge's answer as possible. This type of behavior would damage the validity of the results. However, since no participant consequently gave answers above or below the advice, this is unlikely to have been a problem. A participant could also realize the bias of the system and change their answer opposite to the advice, without the intention of gaming the system. The present research setup does not allow us to distinguish between the two. Other interpretations of WOA values outside the 0-1 range are participants changing their minds after rereading the case description or typing mistakes.

Future research

The goal of the present study was to see whether value similarity was at all relevant for values expressed in the form of AI bias. Therefore, only two oppositely biased advice conditions were included. A future study could incorporate an unbiased AI. Such a study could examine whether value similarity leads to the formation of appropriate trust or disuse instead. It could be problematic if congruence between a person's values and advice results in them trusting biased advice more than unbiased advice. It would also be interesting to compare the adoption rates between a value similar biased AI and an unbiased AI. If minor adjustments of AI recommendations to better align with user values can indeed increase adoption rates, then the overall collaborative process between humans and AI may be improved. Furthermore, such a study may opt to include participants with more moderate beliefs. The present study tried to select participants with clear beliefs regarding prison sentencing. It is unknown if the observed influence of value similarity on trust and compliance will generalize to people with more moderate beliefs.

As discussed, the present research does not use a single significant drop in accuracy but instead works with an overall biased AI that randomly makes mistakes. The results show that while value similarity increases base trust, inaccuracies still strongly negatively affect trust. Future research could investigate if a single large drop in accuracy in an otherwise accurate system is influenced by value similarity. It may be the case that there is an accuracy threshold after which AI error or bias is no longer viewed as a value judgment and is solely seen as a mistake. For instance, the present study provides evidence that a lenient person is more trusting of an AI that provides lenient sentences. However, it may be that if that lenience becomes too great (e.g., less than half the sentence of a judge) this would no longer be perceived as lenience but instead as a mistake, resulting in no value similarity.

Results of the present study show mind perception to be a significant predictor of value similarity with an AI. This provides some evidence that people made value attributions to AI as an entity in and of itself. That said, these results cannot make any detailed inferences about the source of value perceptions of AI. The question remains whether values are perceived as an inherent characteristic of the AI, or as the values of its developers. Future research is required to gain a better understanding of value attribution to AI.

Conclusion

The primary goal of this study was to examine the effect of value similarity on trust in biased AI advice. Additionally, differences in trust between AI and human advisors, trust development over time, and mind perception of AI were analyzed. The results show that people trust biased advice more if they perceive an AI as having similar values. These results add to the existing AI value similarity literature by showing that values implied through AI advice in the form of bias can foster a sense of value similarity. This suggests that values play a role in trust formation with AI even when a system is not explicitly designed to express values. Furthermore, the results indicate that trust resulting from value similarity translates into compliance with AI advice. These findings show the potential benefits of incorporating user values into AI design. The adoption of assistive AI may be improved by slightly adjusting AI advice to align with user values. However, this study also shows that value alignment could lead to over trust in biased systems. Greater compliance with advice due to value similarity could lead to disuse. The possibility that users assess advice from AI based on its alignment with their personal values should be considered when developing decision-making aids. Especially when these aids will be used in emotional or ethical situations where choices are partly based on personal opinions, beliefs, or values.

References

Adriaenssen, A., & Aertsen, I. (2015). Punitive attitudes: Towards an operationalization to measure individual punitivity in a multidimensional way. European Journal of Criminology, 12(1), 92-112. https://doi.org/10.1177/1477370814535376

Ala-Pietilä, P., Bonnet, Y., Bergmann, U., Bielikova, M., Bonefeld-Dahl, C., Bauer, W., ... & Van Wynsberghe, A. (2020). The assessment list for trustworthy artificial intelligence (ALTAI). European Commission.

Beilmann, M., & Lilleoja, L. (2015). Social trust and value similarity: The relationship between social trust and human values in Europe. Studies of transition states and societies, 7(2), 19-30.

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

Burton, J. W., Stein, M. K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. Journal of Behavioral Decision Making, 33(2), 220-239.
https://doi.org/10.1002/bdm.2155

Cabiddu, F., Moi, L., Patriotta, G., & Allen, D. G. (2022). Why do users trust algorithms? A review and conceptualization of initial trust and trust over time. European Management Journal.
https://doi.org/10.1016/j.emj.2022.06.001

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. Journal of Marketing Research, 56(5), 809-825.
https://doi.org/10.1177/0022243719851788

Cazier, J. A., Shao, B. B., & Louis, R. D. S. (2006). E-business differentiation through value-based trust. Information & Management, 43(6), 718-727.
https://doi.org/10.1016/j.im.2006.03.006

Chancey, E. T., Bliss, J. P., Proaps, A. B., & Madhavan, P. (2015). The role of trust as a mediator between system characteristics and response behaviors. Human factors, 57(6), 947-958.
DOI: 10.1177/0018720815582261

Cvetkovich, G. (2013). The attribution of social trust. In Social trust and the management of risk (pp. 67-75). Routledge.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. International journal of human-computer studies, 58(6), 697-718.
https://doi.org/10.1016/S1071-5819(03)00038-7

Earle, T. C., & Cvetkovich, G. (1995). Social trust: Toward a cosmopolitan society. Greenwood Publishing Group.

Feng, S., & Boyd-Graber, J. (2019). What can ai do for me? evaluating machine learning interpretations in cooperative play. In Proceedings of the 24th International Conference on Intelligent User Interfaces (pp. 229-239).
https://doi.org/10.1145/3301275.3302265

Gigliotti, L. M., Sweikert, L. A., Cornicelli, L., & Fulton, D. C. (2020). Minnesota landowners' trust in their department of natural resources, salient values similarity and wildlife value orientations. Environment Systems and Decisions, 40(4), 577-587.
https://doi.org/10.1007/s10669-020-09766-z

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 14(2), 627-660.
https://doi.org/10.5465/annals.2018.0057

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. Journal of Artificial Intelligence Research, 62, 729-754.
https://doi.org/10.1613/jair.1.11222

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. science, 315(5812), 619-619.
DOI: 10.1126/science.1134475

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. Psychological inquiry, 23(2), 101-124.
https://doi.org/10.1080/1047840X.2012.651387

Hafenbrädl, S., Waeger, D., Marewski, J. N., & Gigerenzer, G. (2016). Applied decision making with fast-and-frugal heuristics. Journal of Applied Research in Memory and Cognition, 5(2), 215-231.
https://doi.org/10.1016/j.jarmac.2016.04.011

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. Organizational behavior and human decision processes, 70(2), 117-133.
https://doi.org/10.1006/obhd.1997.2697

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2022). lab. js: A free, open, online study builder. Behavior Research Methods, 54(2), 556-573.
https://doi.org/10.3758/s13428-019-01283-5

Himmelstein, M. (2022). Decline, adopt or compromise? A dual hurdle model for advice utilization. Journal of Mathematical Psychology, 110, 102695.
https://doi.org/10.1016/j.jmp.2022.102695

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human factors, 57(3), 407-434.
DOI: 10.1177/0018720814547570

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608.
https://doi.org/10.48550/arXiv.1812.04608

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural equation modeling: a multidisciplinary journal, 6(1), 1-55.
https://doi.org/10.1080/10705519909540118

Ives, C. D., & Kendal, D. (2014). The role of social values in the management of ecological systems. Journal of environmental management, 144, 67-72.
https://doi.org/10.1016/j.jenvman.2014.05.013

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. European Journal of Information Systems, 31(3), 388-409.
https://doi.org/10.1080/0960085X.2021.1927212

Langer, M., König, C. J., Back, C., & Hemsing, V. (2022). Trust in Artificial Intelligence: Comparing trust processes between human and automated trustees in light of unfair bias. Journal of Business and Psychology, 1-16.
https://doi.org/10.1007/s10869-022-09829-9

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human factors, 46(1), 50-80.

Li, Z., Terfurth, L., Woller, J. P., & Wiese, E. (2022). [PREPRINT] Mind the Machines: Applying Implicit Measures of Mind Perception to Social Robotics.
Preprint DOI: 10.31234/osf.io/2ezfj

Liu, Y., & Moore, A. (2022). A Bayesian Multilevel Analysis of Belief Alignment Effect Predicting Human Moral Intuitions of Artificial Intelligence Judgements. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44, No. 44).

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

Lockey, S., Gillespie, N., Holm, D., & Someh, I. A. (2021). A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions.
https://doi.org/10.24251/hicss.2021.664

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. In Proceedings of the 11th Australasian Conference on Information Systems, 6–8.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35.
https://doi.org/10.1145/3457607

Mehrotra, S., Jonker, C. M., & Tielman, M. L. (2021). More Similar Values, More Trust?-the Effect of Value Similarity on Trust in Human-Agent Interaction. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 777-783).
https://doi.org/10.1145/3461702.3462576

Nothdurft, F., Richter, F., & Minker, W. (2014). Probabilistic human-computer trust handling. In Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL) (pp. 51-59).

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

Pauketat, J. V., & Anthis, J. R. (2022). Predicting the moral consideration of artificial intelligences. Computers in Human Behavior, 136, 107372.
https://doi.org/10.1007/s43681-023-00260-1

Pharmer, R. L., Wickens, C. D., Clegg, B. A., & Smith, C. A. P. (2021). Effect of Procedural Elements on Trust and Compliance with anImperfect Decision Aid. In Proceedings of the Human Factors and

Ergonomics Society Annual Meeting (Vol. 65, No. 1, pp. 633-637). Sage CA: Los Angeles, CA: SAGE Publications.

Poortinga, W., & Pidgeon, N. F. (2006). Prior attitudes, salient value similarity, and dimensionality: Toward an integrative model of trust in risk regulation 1. Journal of Applied Social Psychology, 36(7), 1674-1700.
https://doi.org/10.1111/j.0021-9029.2006.00076.x

Prahl, A., & Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted?. Journal of Forecasting, 36(6), 691-702.
https://doi.org/10.1002/for.2464

Renier, L. A., Mast, M. S., & 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

Riedl, R. (2022). Is trust in artificial intelligence systems related to user personality? Review of empirical evidence and future research directions. Electronic Markets, 1-31.
https://doi.org/10.1007/s12525-022-00594-4

Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. Human factors, 58(3), 377-400.
doi: 10.1177/0018720816634228

Schwartz, S. H. (1994). Are there universal aspects in the structure and contents of human values?. Journal of social issues, 50(4), 19-45.
https://doi.org/10.1111/j.1540-4560.1994.tb01196.x

Shank, D. B., North, M., Arnold, C., & Gamez, P. (2021). Can mind perception explain virtuous character judgments of artificial intelligence?. Technology, Mind, and Behavior, 2(3).
https://doi.org/10.1037/tmb0000047

Siegrist, M., Cvetkovich, G., & Roth, C. (2000). Salient value similarity, social trust, and risk/benefit perception. Risk analysis, 20(3), 353-362.
https://doi.org/10.1111/0272-4332.203034

Sitkin, S. B., & Roth, N. L. (1993). Explaining the limited effectiveness of legalistic "remedies" for trust/distrust. Organization science, 4(3), 367-392.
https://doi.org/10.1287/orsc.4.3.367

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making?. International Journal of Human-Computer Studies, 51(5), 991-1006.
https://doi.org/10.1006/ijhc.1999.0252

Spiranovic, C. A., Roberts, L. D., & Indermaur, D. (2012). What predicts punitiveness? An examination of predictors of punitive attitudes towards offenders in Australia. Psychiatry, Psychology and Law, 19(2), 249-261.
https://doi.org/10.1080/13218719.2011.561766

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. American journal of political science, 50(3), 755-769.
https://doi.org/10.1111/j.1540-5907.2006.00214.x

Tolmeijer, S., Christen, M., Kandul, S., Kneer, M., & Bernstein, A. (2022). Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making. In CHI Conference on Human Factors in Computing Systems (pp. 1-17).
https://doi.org/10.1145/3491102.3517732

Twyman, M., Harvey, N., & Harries, C. (2008). Trust in motives, trust in competence: Separate factors determining the effectiveness of risk communication. Judgment and Decision Making, 3(1), 111.
https://doi.org/10.1017/S1930297500000218

Verberne, F. M., Ham, J., & Midden, C. J. (2012). Trust in smart systems: Sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars. Human factors, 54(5), 799-810.
https://doi.org/10.1177/0018720812443825

Vereschak, O., Bailly, G., & Caramiaux, B. (2021). How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1-39.
https://doi.org/10.1145/3476068

Yaniv, I. (2004). The benefit of additional opinions. Current directions in psychological science, 13(2), 75-78.
https://doi.org/10.1111/j.0963-7214.2004.00278.x

Yokoi, R., & Nakayachi, K. (2019). The effect of shared investing strategy on trust in artificial intelligence. The Japanese Journal of Experimental Social Psychology, 59(1), 46-50.
https://doi.org/10.2130/jjesp.1819

Yokoi, R., & Nakayachi, K. (2021a). Trust in autonomous cars: exploring the role of shared moral values, reasoning, and emotion in safety-critical decisions. Human factors, 63(8), 1465-1484.
DOI:10.1177/0018720820933041

Yokoi, R., & Nakayachi, K. (2021b). The effect of value similarity on trust in the automation systems: A case of transportation and medical care. International Journal of Human–Computer Interaction, 37(13), 1269-1282.
https://doi.org/10.1080/10447318.2021.1876360

Yokoi, R., Eguchi, Y., Fujita, T., & Nakayachi, K. (2020). Artificial intelligence is trusted less than a doctor in medical treatment decisions: Influence of perceived care and value similarity. International Journal of Human–Computer Interaction, 37(10), 981-990.
https://doi.org/10.1080/10447318.2020.1861763

Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., & Chen, F. (2017). User trust dynamics: An investigation driven by differences in system performance. In Proceedings of the 22nd international conference on intelligent user interfaces (pp. 307-317).
https://doi.org/10.1145/3025171.3025219

Appendix

Measurement items

Final trust

11-item trust scale adapted from Madsen & Gregor (2000).

To what extent do you agree with the following statements.

1. The legal AI/Expert always provides the advice I require to make my decision.
2. The legal AI/Expert performs reliably.
3. I can rely on the legal AI/Expert to give advice properly.
4. The legal AI/Expert analyzes cases consistently.
5. The legal AI/Expert uses appropriate methods to reach decisions.
6. The legal AI/Expert has sound knowledge about these types of legal cases.
7. The legal AI/Expert makes use of all the knowledge and information available to form the advice.
8. I believe advice from the legal AI/Expert even when I don't know for certain that it is correct.
9. If I am not sure about a decision, I have faith that the legal AI/Expert will provide the best solution.
10. When the legal AI/Expert gives unusual advice, I am confident that the advice is correct.
11. Even if I have no reason to expect the legal AI/Expert will be able to solve a difficult case, I still feel certain about the advice.

Case trust

3-item trust scale adapted from Madsen & Gregor (2000).

To what extent do / are you...

1. have faith in the legal AI/Expert's advice?
2. believe the legal AI/Expert's advice more than yourself?
3. confident the legal AI/Expert provides the best solution?

Value similarity

Semantic differentials taken from Siegrist et al. (2000).

Please indicate in the following pairs of opposite words how similar or dissimilar you are compared to the legal AI/Expert you were getting advice from during the previous tasks.

Same values - Different values
Acts as I would - Acts different than I would
Same goals - Different goals
Thinks like me - Thinks unlike me
Same opinions - Different opinions

Punitive attitudes

Punitive attitudes scale taken from Spiranovic, Roberts & Indermaur (2012).

Please indicate to what extent you agree with the following statements regarding prison sentencing and crime.

1. The death penalty should be the punishment for murder.
2. People who break the law should be given stiffer sentences.
3. The courts are too soft on offenders.
4. The tougher the sentence, the less likely an offender is to commit more crime.
5. Rehabilitation is not taken seriously by criminals.

6. High crime rates are mainly an indication or sign that punishments are not severe enough.
7. The most effective response to crime is to have harsher sentences.

Mind perception
Agentic mind perception scale adapted from Li et al. (2022).

Please indicate to what extent you agree with the following statements. The legal AI/Expert can:
1. ...have thoughts
2. ...have beliefs
3. ...tell right from wrong
4. ...plan future actions
5. ...understand others' minds
6. ...set goals
7. ...exercise self-control
8. ...uphold moral values

Case descriptions

Full case descriptions and the judge's verdict in months of jail time. The case order was randomized for each participant (except for the practice case).

Case 1, judge's verdict = 48
The defendant entered a supermarket with the intention of robbing it. After completing the robbery, while leaving the store, the defendant was stopped by two people, the later victims. The defendant pushed the victims onto the ground and stabbed them multiple times with a knife. One victim was stabbed in the arm and the other in the stomach. The defendant also threatened to murder the victims. In addition to the supermarket robbery, the defendant extorted a third victim. The defendant threatened that victim by showing them the defendant's knife. The Public Prosecution Office concluded that the defendant was guilty of attempted murder, robbery, and extortion.

Case 2, judge's verdict = 36
The defendant, together with one co-defendant, forcibly gained access to a house. The co-defendant had already prepared the crime the day before the robbery by turning off the security cameras and opening the garage door through which the defendant entered the house. The defendant used force against the resident of the house by pushing them to the ground and hitting them. In addition, the defendant threatened to murder the resident. When the resident fiercely resisted, the defendant fled. The defendant stole a wallet containing money and bank cards. The Public Prosecution Office concluded that the defendant was guilty of residential burglary.

Case 3, judge's verdict = 54
After a heated argument between the defendant and the victim, the defendant attempted to kill the victim. The defendant stabbed and struck the victim with a knife. The defendant also threatened to stab the victim in the mouth. The victim survived the attack. The victim did not suffer life-threatening injuries, however, they do suffer from PTSD as a result of the crime. The Public Prosecution Office concluded that the defendant was guilty of attempted murder.

Case 4, judge's verdict = 42
The defendant, alongside two confederates, entered a fish store and told the two store owners that they must pay £4000 per month to the defendant or else the store would be destroyed. The

defendant threatened to use violence against the store owners if the money was not paid. The defendant intimidated the store owners by pretending to have a weapon. Additionally, the defendant hit both victims repeatedly on their heads and in their stomachs. The defendant has previously served jail time for similar violent crimes. The Public Prosecution Office concluded that the defendant was guilty of extortion and assault.

Case 5, judge's verdict = 16
The defendant used phishing servers, sites and software with the aim of committing fraud. By posing as an employee of a banking service, the defendant tried to obtain the personal data of customers by making them click on a link. Furthermore, the defendant laundered money for a total of £90,000. In addition to these crimes, the defendant forged Covid-19 travel certificates. During the search of the defendant’s apartment, a fake realistic-looking firearm was found. The Public Prosecution Office concluded that the defendant was guilty of computer intrusion, sending spam, money laundry and forgery.

Case 6, judge's verdict = 24
The defendant entered an office building with the intention to rob it. The defendant concealed their face and was armed with a gun. They pointed the gun at the people in the building and threatened to shoot them if they did not cooperate. While in the building, the defendant grabbed and dragged people and held the gun close to their heads and necks. Throughout the robbery, the defendant continuously threatened to shoot the office employees. The defendant stole a total of £60.000 and a car. The Public Prosecution Office concluded that the defendant was guilty of armed robbery.

Case 7, judge's verdict = 48
The defendant, who has a psychiatric disability, stabbed their romantic partner in the chest with a knife after they had gotten into a verbal conflict. As a result, the victim went into cardiac arrest. The defendant alerted emergency services and attempted to resuscitate the victim. The victim underwent surgery and spent months in the hospital. The victim suffered brain damage and doctors do not expect them to make a full recovery. The Public Prosecution Office concluded that the defendant was guilty of attempted murder.

Case 8, judge's verdict = 60
The defendant was driving a car without owning a driver's license. The defendant was also under the influence of alcohol and cannabis while driving. There were two other passengers in the car. During this drive, the defendant significantly exceeded the speed limit. This led to an accident in which the vehicle crashed into a tree on the side of the road. One of the passengers was killed in this accident, and the other was seriously injured. Prior to this incident, the defendant already had a run-in with the police for driving without a license. The Public Prosecution Office concluded that the defendant was guilty of reckless driving resulting in death.

Case 9, judge's verdict = 32
The defendant attempted to commit two robberies. First, they entered a supermarket while concealing their face. They grabbed a cashier by the arm and demanded that they hand over the money from the cash register while threatening them with a knife. Second, they entered a café, looking for a person they had argued with. Upon not finding this person inside, they then decided to rob the café while pretending to have a concealed firearm. In both cases, the defendant was unsuccessful and no money was taken. The Public Prosecution Office concluded that the defendant was guilty of multiple counts of attempted armed robbery.

Case 10, judge's verdict = 24
The defendant drove their car to a psychiatric detention facility to break out two inmates. Using an angle grinder, the defendant created a hole in the fence surrounding the facility and cut through the power cables of various security systems. This destroyed both the fence and the security systems. After this, the defendant collected the two inmates and drove them to separate locations. The Public Prosecution Office concluded that the defendant was guilty of breaking out two inmates from a psychiatric detention facility.

Case 11, judge's verdict = 30
The defendant transported 5kg of heroin from The UK to Germany. Given the large quantity of heroin, the Public Prosecution Office finds it reasonable to assume that the heroin was intended for further distribution. Given the significant negative societal impact of heroin, the court found this fact to weigh heavily on the severity of the offence. The Public Prosecution Office concluded that the defendant was guilty of exporting heroin.

Case 12, judge's verdict = 60
In ten individual cases, the defendant entered the homes of elderly citizens. The defendant pretended to be a home cleaner. Inside the victims' homes, the defendant stole money (up to £5,000), jewellery and bank cards. If a bank card was stolen, the defendant subsequently called the victim and pretended to be a bank employee. This way they acquired the victim's card number, which they then used to withdraw money from the victim's account. The Public Prosecution Office concluded that the defendant was guilty of multiple counts of fraud.

Case 13, judge's verdict = 20
In a state of intoxication, the defendant, together with other persons, repeatedly attacked a person in a public place. The defendant allegedly played a leading role in this. The defendant attacked the victim several times until they fell to the ground. While the victim was on the ground, the defendant repeatedly punched and kicked the victim’s head. Afterwards, the defendant left the victim on the ground without alerting emergency services. The court found that the defendant’s actions could have led to the victim’s death. The Public Prosecution Office concluded that the defendant was guilty of complicity in attempted manslaughter.

Case 14, judge's verdict = 24
The defendant entered a gas station store and instructed an employee to open the cash register while threatening them with a sharp object. After this, the defendant grabbed the money from the register and threw the coins in the face of the employee. In addition to this robbery, the defendant verbally and physically threatened a security guard at the addiction rehabilitation center where the defendant was a resident. The Public Prosecution Office concluded that the defendant was guilty of robbery and verbal and physical threats.

Practice case, judge's verdict = 24
After a night out, the defendant got into an altercation with a stranger on the street. During this, the defendant stabbed the victim with a knife in the lower-left abdomen. While the victim survived, if the knife had punctured the abdominal wall of the victim, it could have had fatal consequences. The defendant showed a lack of respect for human life and violated the victim's physical integrity with their act. The Public Prosecution Office concluded that the defendant was guilty of attempted manslaughter.

Factor loadings

Item	Psychological trust	Mind perception	Punitive attitudes	Value similarity
Trust item 1	0.72	0.17	-0.02	0.35
Trust item 2	0.86	0.09	-0.09	0.21
Trust item 3	0.88	0.22	-0.06	0.17
Trust item 4	0.70	0.13	0.09	0.17
Trust item 5	0.76	0.19	0.08	0.24
Trust item 6	0.78	0.23	0.01	0.12
Trust item 7	0.75	0.1	0.04	0.15
Trust item 8	0.81	0.08	0.06	0.19
Trust item 9	0.79	0.16	0.04	0.20
Trust item 10	0.78	0.19	0.03	0.18
Trust item 11	0.78	0.19	0.16	0.24
Value similarity item 1	0.34	0.15	0.01	0.71
Value similarity item 2	0.34	0.14	-0.03	0.76
Value similarity item 3	0.35	0.04	0.09	0.57
Value similarity item 4	0.25	0.14	0.02	0.84
Value similarity item 5	0.33	0.08	0.08	0.81
Mind perception item 1	0.04	0.93	0.04	0.03
Mind perception item 2	0.04	0.88	0.09	0.06
Mind perception item 3	0.23	0.78	0.23	0.19
Mind perception item 4	0.21	0.61	0.14	0.06
Mind perception item 5	0.21	0.82	0.11	0.11
Mind perception item 6	0.26	0.64	0.11	0.08
Mind perception item 7	0.14	0.75	0.12	0.05
Mind perception item 8	0.28	0.76	0.17	0.12
Punitive attitudes item 1	-0.06	0.10	0.63	0.04
Punitive attitudes item 2	0.15	0.13	0.83	-0.02
Punitive attitudes item 3	0.09	0.16	0.85	-0.08
Punitive attitudes item 4	-0.07	0.17	0.66	0.05
Punitive attitudes item 5	0.11	-0.03	0.68	0.03
Punitive attitudes item 6	-0.03	0.15	0.89	0.03
Punitive attitudes item 7	0.03	0.15	0.90	0.10
SS loadings	7.64	5.35	4.51	3.37
Proportion of Variance	0.25	0.17	0.15	0.11
Cumulative Variance	0.25	0.42	0.56	0.67

Table 12. Factor loadings of all measured variables.