Human Perceptions on Moral Responsibility of AI: A Case Study in AI-Assisted Bail Decision-Making

Gabriel Lima
gabriel.lima@kaist.ac.kr
School of Computing, KAIST
Data Science Group, IBS
Republic of Korea

Nina Grgić-Hlača
nghlaca@mpi-sws.org
Max Planck Institute for Software Systems
Max Planck Institute for Research on Collective Goods
Germany

Meeyoung Cha
mcha@ibs.re.kr
Data Science Group, IBS
School of Computing, KAIST
Republic of Korea

ABSTRACT

How to attribute responsibility for autonomous artificial intelligence (AI) systems' actions has been widely debated across the humanities and social science disciplines. This work presents two experiments (N=200 each) that measure people's perceptions of eight different notions of moral responsibility concerning AI and human agents in the context of bail decision-making. Using real-life adapted vignettes, our experiments show that AI agents are held causally responsible and blamed similarly to human agents for an identical task. However, there was a meaningful difference in how people perceived these agents' moral responsibility; human agents were ascribed to a higher degree of present-looking and forward-looking notions of responsibility than AI agents. We also found that people expect both AI and human decision-makers and advisors to justify their decisions regardless of their nature. We discuss policy and HCI implications of these findings, such as the need for explainable AI in high-stakes scenarios.

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in HCI*; • **Applied computing** → **Psychology**; *Law*.

KEYWORDS

AI, Moral Responsibility, Responsibility, Moral Judgment, Blame, Liability, COMPAS, Bail Decision-Making

ACM Reference Format:
Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. 2021. Human Perceptions on Moral Responsibility of AI: A Case Study in AI-Assisted Bail Decision-Making. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3411764.3445260

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445260

1 INTRODUCTION

Who should be held responsible for the harm caused by artificial intelligence (AI)? This question has been debated for over a decade since Matthias' landmark essay on the *responsibility gap* of autonomous machines [68]. This gap is posed by highly autonomous and self-learning AI systems. Until now, scholars in multiple disciplines, including ethics, philosophy, computer science, and law, have suggested possible solutions to this moral and legal dilemma. Optimistic views proclaim that the gap can be bridged by proactive attitudes of AI designers, who should readily take responsibility for any harm [20, 72]. Some even propose to hold AI systems responsible per se [91], viewing human-AI collaborations as extended agencies [45, 48]. In contrast, pessimistic views question whether this gap can be bridged at all, since there might not exist appropriate subjects of retributive blame [26] nor it makes sense to hold inanimate and non-conscious entities responsible for their actions [16, 89, 96].

Most research on the responsibility gap has been normative in that they prescribed ethical principles and proposed solutions. However, there is a growing need for practical and proactive guidelines; as Mittelstadt puts it, "principles alone cannot guarantee ethical AI" [69]. Some even argue that normative approaches are inappropriate as they can hurt AI's adoption in the long run [12]. In contrast, relatively little attention has been paid to understanding the public's views on this issue, who are likely the most affected stakeholder when AI systems are deployed [78].

We conducted two survey studies (N=200 each) that collect the public perception on moral responsibility of AI and human agents in high-stakes scenarios. We approached the pluralistic view of responsibility and considered eight distinct notions compiled from philosophy and psychology literature. Real-life adapted vignettes of AI-assisted bail decisions were used to observe how people attributed specific meanings of responsibility to *i*) AI advisors vs. human advisors and *ii*) AI decision-makers vs. human decision-makers. Our study employed a within-subjects design where all participants were exposed to a diverse set of vignettes addressing distinct possible outcomes from bail decisions.

Our findings suggest that the eight notions of responsibility considered can be re-grouped into two clusters: one encompasses *present-looking and forward-looking* notions (e.g., responsibility-as-task, as-power, as-authority, as-obligation), and the other includes *backward-looking* notions (e.g., blame, praise, liability) and *causal*

determinations. We discuss how theories of moral responsibility can explain these clusters.

In comparing AI agents against human agents, we found a striking difference in the way people attribute responsibility. A substantially higher degree of the present- and forward-looking notions were attributed to human agents than AI agents. This means that AI agents were assigned the responsibility to complete and oversee the same task to a lesser extent than human agents. No difference, however, was observed for the backward-looking responsibility notions. This finding suggests that blame, liability, and causal responsibility were ascribed equally to AI and human agents, despite electronic agents not being appropriate subjects of liability and blame [16, 26, 89]. In addition to these findings, we found that people expect both human and AI agents to justify their decisions.

The findings of this study have several implications for the development and regulation of AI. Using the proposition of morality as a human-made social construct that aims to fulfill specific goals [91, 93], we highlight the importance of users and designers taking responsibility for their systems while being held responsible for any norm-violating outcomes. We also discuss the possibility of holding AI systems responsible per se [61] *alongside* other human agents, as a possible approach congruent to the public opinion.

2 BACKGROUND

2.1 Theories of (Moral) Responsibility

Theories of moral responsibility date back to Aristotle, who argued that an entity should satisfy both freedom and epistemic conditions to appropriately be ascribed to moral responsibility. Agents must act freely, without coercion, and understand their actions. Although recent scholarly work does not directly challenge these Aristotelian conditions, they argue that moral responsibility cannot be explained as a single concept, but that it involves a relatively pluralistic definition of what it means to hold someone morally responsible [87, 102].

Scanlon [85] has proposed moral responsibility to be a bipartite concept. One is that there is an account of *being* responsible in rendering an agent worthy of moral appraisal. Another is that it is also possible to *hold* one responsible for specific actions and consequences. Expanding this bipartite concept, Shoemaker [87] has proposed three different concepts of moral responsibility: attributability, answerability, and accountability. Various other definitions have been proposed [102], including structured notions of what responsibility is [104] and how they are connected [14, 34].

Attributing responsibility to an entity can be both descriptive (e.g., causal responsibility) and normative (e.g., blameworthiness). For the former, one might ask if an agent *is* responsible for an action or state-of-affairs, while the latter concerns whether one *should* attribute responsibility to an agent. Responsibility can also be divided into backward-looking notions if they evaluate a past action and possibly lead to reactive attitudes [106], or forward-looking notions if they prescribe obligations.

Responsibility can take many forms. It not only addresses the moral dimension of society but also tackles legal concepts and other descriptive notions. One can be held legally responsible (i.e., liable) regardless of their moral responsibility, as in the case of strict or vicarious liability. Stating that an agent is causally responsible for a state-of-affairs does not necessarily prescribe a moral evaluation of the action.

Holding an agent "responsible" fulfills a wide range of social and legal functions. Legal scholars state that punishment (which could be seen as a form of holding an agent responsible, e.g., under criminal liability) aims to reform the wrongdoers, deter re-offenses and similar actions, and resolve retributive sentiments [4, 99]. Previous work has addressed how and why people assign responsibility to various agents. The general public might choose to hold a wrongdoer responsible for restoring moral coherence [22] or reaffirming a communal moral values [109]. Psychological research indicates that people base much of their responsibility attribution on retributive sentiments rather than deterrence [18], while overestimating utilitarian goals in their ascription of punishment (i.e., responsibility) [17]. Intentionality also determines how much responsibility is assigned to an entity [70]; people look for an intentional agent to hold responsible and infer other entities' intentionality upon failure to find one [40].

2.2 Techno-Responsibility Gaps

AI systems and robots are being widely adopted across society. Algorithms are used to choose which candidate is most fit for a job position [111], decide which defendants are granted bail [33], guide health-related decision [73], and assess credit risk [49]. AI systems are often embedded into robots or machines, such as autonomous vehicles [13] and robot soldiers [3]. A natural question here is: if an AI system or a robot causes harm, who should be held responsible for their actions and consequences?

In answering this question, some scholars have defended the existence of a (techno-)responsibility gap [68] for autonomous and self-learning systems.[1] The autonomous component of AI and robots challenges the control condition of responsibility attribution. Simultaneously, their self-learning capabilities and opacity do not allow users, designers, and manufacturers to foresee consequences. Similarly to the "problem of many hands" in the assignment of responsibility to collective agents [102], AI and robots suffer from the "problem of many things," i.e., current systems are composed of various interacting entities and technologies, making the search for a responsible entity harder [24]. Scholars have extensively discussed the assignment of responsibility for autonomous machines' actions and have expanded this gap to more specific notions of responsibility [5, 8, 54] and its functions [26, 62].

Although a clear separation is fuzzy, one may find two schools of thought on the responsibility gap issue. One side argues that designers and manufacturers should take responsibility for any harm caused by their "tools." [16, 31] Supervisors and users of these systems should also take responsibility for their deployment, particularly in consequential environments like the military as argued by Champagne and Tonkens [20]. The exercise of agency by these systems can be viewed as a human-robot collaboration, in which humans supervise and manage the agency of AI and robots [72]. Humans should focus on their relationship to the patients of their responsibility to answer for the actions of autonomous systems [24].

[1] Scholars also raise doubt on the existence of techno-responsibility gaps, arguing that moral institutions are dynamic and flexible and can deal with these new technological artifacts [53, 95].

Likewise, other authors argue that society should hold humans responsible because doing so for a machine would be meaningless as it does not understand the consequences of their actions or the reactive attitudes towards them [89, 96], possibly undermining the definition of responsibility [47].

On the opposite side, some scholars propose autonomous systems could be held responsible per se [61]. From a legal perspective, non-human entities (e.g., corporations) can be held responsible for any damage that they may cause [103]. These scholars often view these human-AI collaborations as extended agencies where all entities should be held jointly responsible [45, 48]. AI and robots are part of the socio-technological ensemble, in which responsibility can be distributed across multiple entities with varying degrees [32]. These proposals arguably contribute to legal coherence [98], although it could also lead to various repercussions in moral and legal institutions [8]. Empirical findings indicate that people attribute responsibility to these systems [7, 62], although to a lesser extent than human agents. According to some scholars, holding AI and robots responsible per se could fulfill specific social goals [23] and promote critical social functions [11, 91].

The regulation of AI and robots poses new challenges to policymaking, as in the previously introduced techno-responsibility gap, which society must discuss at large [24]. The "algorithmic social contract" requires inputs from various stakeholders, whose opinion should be weighed for the holistic crafting of regulations [78]. It is crucial to understand how people perceive these systems before their wide deployment [80]. Our responsibility practices depend on folk-psychology [15] (i.e., how people perceive the agents involved in social practices [91]). Literature exists on the public perception of moral and legal issues concerning AI [6, 7, 62]. However, little data-driven research has collected public opinion on how responsibility should be attributed for AI and robots' actions.

2.3 Responsibility, Fairness, Trust in HCI Literature

A growing number of HCI research has been devoted to understanding how people perceive algorithmic decisions and their consequences in society. For instance, Lee et al. studied people's perception of trust, fairness, and justice in the context of algorithmic decision-making [56, 57] and proposed how to embed these views into a policymaking framework [58]. Other scholars explored people's perceptions of procedural [41] and distributive [84, 90] aspects of algorithmic fairness and studied how they relate to individual differences [42, 76, 108]. Nonetheless, little attention is paid to the public attribution of (moral) responsibility to stakeholders (e.g., [43, 56, 81]), particularly the prospect of responsibility ascription to the AI system per se. The current study contributes by addressing the public perception of algorithmic decision-making through the lens of moral responsibility.

Existing studies addressing how users might attribute blame to automated agents have mostly focused on robots. For instance, Malle et al. observed that people's moral judgments between human and robotic agents differed in that respondents blamed robots to a more considerable extent had they not taken a utilitarian action [67]. Furlough et al. found that respondents attributed similar levels of blame to robotic agents and humans when robots were described as autonomous and at the same time the leading cause of harm [37]. However, these studies and many others [52, 59, 105] tackle a singular notion of responsibility related to blameworthiness [102]. The present research explores multiple notions of moral responsibility of both human and AI agents involved in decision-making.

3 METHODOLOGY

3.1 Algorithmic Decision-Making

AI-based algorithms are now used to assist humans in various scenarios, including high-stakes tasks such as medical diagnostics [35] and bail decisions [2]. These algorithms do not make decisions themselves, but rather "advise" humans in their decision-making processes. One such algorithm is the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) tool, used by the judicial system in the US to assist bail decisions and sentencing [2]. Several studies have analyzed the fairness and bias aspects of this risk assessment algorithm, e.g., [9, 33, 43].

This study makes use of publicly available COMPAS data released by ProPublica [2] and considers the machine judgments as either an AI advisor (later in Study 1) or an AI decision-maker (in Study 2). As stimulus material, we use real-world data obtained from a previous analysis of the tool [2], which focused on its application in bail decision-making. This dataset contains information about 7,214 defendants subjected to COMPAS screening in Broward County, Florida, between 2013 and 2014.

We use 100 randomly selected cases from this dataset, the corresponding bail suggestions, and information about whether the defendant re-offended within two years of sentencing. The sampled data was balanced concerning these variables. Each defendant's COMPAS score ranges from 1 to 10, with ten indicating the highest risk of re-offense or nonappearance in court. In this study, scores 1 to 5 were labeled "grant bail" and 6 to 10 were labeled "deny bail."

3.2 The Plurality of Responsibility

Ascribing responsibility is a complex moral and legal practice that encompasses various functions, entities, and social practices [71, 91]. Responsibility has multiple distinct meanings depending on its purpose and requirements. The current study revisits eight notions of responsibility compiled from psychology and philosophy. All of these notions originated from Van de Poel's work [101, 102], except for responsibility-as-authority and as-power, which comes from Davis's discussion on professional responsibility [27]. We complement these notions with a wide range of literature ranging from philosophical theories of moral responsibility (e.g., [86, 87]) to approaches in the context of AI systems (e.g., [24, 96]). Although not exhaustive (e.g., we have not addressed virtue-based notions of responsibility as they cannot be easily adapted to AI systems), we highlight how our work differs from previous HCI approaches.

- **Responsibility-as-obligation:**
  E.g., "The (agent) should ensure that the rights of the defendant are protected."
  One could be held responsible-as-obligation through consequentialist, deontological, and virtue-based routes [102]. While an entity could be attributed such meaning of responsibility based on pre-determined consequentialist distribution

principles, the latter two routes presuppose the agent's initiative or promise to see to it that a specific state-of-affairs is brought about. This notion differs from responsibility-as-task as it does not imply that one should be the agent to bring about a specific state-of-affairs, but rather indicates that one *should* fulfill its supervisory duties in the process.

- **Responsibility-as-task:**
  "It is the (agent)'s task to protect the defendant's rights."
  This descriptive notion of responsibility ascribes a specific task to an entity. These assignments do not necessarily define a moral obligation per se [101] and are often accompanied by the understanding that an entity *has* to do something by itself [27]. In our experimental design, we highlight the agent's acting role in completing its task.
- **Responsibility-as-authority:**
  "The (agent) has the authority to prevent further offenses."
  To be responsible-as-authority implies that one *is* in charge of a specific action or state-of-affairs. This notion has also been posed as "responsibility-as-office" by Davis [27] in the context of engineers' professional responsibility. An important aspect of responsibility-as-authority is the possibility of delegating other complementing notions, such as responsibility-as-task, to other agents. We address this meaning of responsibility by explicitly indicating that the agent has the authority in bail decisions.
- **Responsibility-as-power:**
  "The (agent) has the skills needed to protect the rights of the defendant."
  If an entity *has* the skills needed to bring about an action or state-of-affairs, one might ascribe it responsibility-as-power [27]. In other words, having the ability, in terms of competency, knowledge, or expertise, might lead to the assignment of this notion of responsibility.
- **Responsibility-as-answerability:**
  "The (agent) should justify their advice."
  This is related to how one's reasons for acting in a specific manner could be seen under moral scrutiny. Shoemaker proposed this notion of moral responsibility as a form of judgment of one's actions grounded in moral evaluations [87]. Davis proposed a similar meaning of responsibility under a different name, responsibility-as-accountability [27], as the responsibility for explaining specific consequences. Coeckelbergh later applied this concept through a relational approach for actions and decisions made using AI [24].
- **Responsibility-as-cause:**
  "The (agent)'s decision led to the prevention of the re-offense."
  This meaning of responsibility has been further discussed depending on the nature of an action's consequences [27], e.g., being causally responsible for a positive state-of-affairs could lead to the ascription of "good-causation." Causality is also an important pre-condition for other normative notions of responsibility, such as blame, as the blurring of a causal connection raises questions on the foreseeability and control of a specific action. [66, 102]
- **Responsibility-as-blame/praise:**
  "The (agent) should be blamed for the violation of the rights of the defendant." / *"The (agent) should be praised for the protection of the rights of the defendant."*
  Blaming an entity for the consequences of their actions has been debated as adopting certain reactive attitudes towards it [106]. Scholars have also argued that to blame someone is to respond to "the impairment of a relationship," [21, 86] especially towards its constitutive standards [87]. Scholars have debated the possibility of ascribing blame to an automated agent and agree that doing so would not be morally appropriate [26, 96]. Regardless of this consensus, previous studies have found that people attribute a similar degree of blame to robotic and human agents under specific conditions (e.g., [37, 67]).
  As an opposite concept of blame, one may consider "praise" as a positive behavioral reinforcement [51] through which one conveys its values and expectations of the agent [28]. Hence, we consider both blame and praise as responsibility notions in this research.
- **Responsibility-as-liability:**
  "The (agent) should compensate those harmed by the re-offense."
  An entity that is ascribed this responsibility should remedy any harm caused by their actions [102]. Rather than dwelling on the discussion addressing the mental states of AI and robots and their arguable incompatibility with criminal law and its assumption of *mens rea* [39, 55, 60], we address this notion from a civil law perspective. Scholars propose 'making victims whole' as the primary goal of tort law [77], and hence, we similarly address responsibility-as-liability. We also add that the idea of holding automated agents liable became prominent after the European Parliament considered adopting a specific legal status for "sophisticated autonomous robots" [29]. Nevertheless, it is important to note that current AI systems cannot compensate those harmed, as they do not possess any assets to be confiscated [16].

3.3 Survey Design

• Study 1: AI as Advisor

To study how the perceived responsibility for bail decisions differs when judges are *advised* by the COMPAS tool or by another human judge, we considered the following scenario:

> *Imagine that you read the following story in your local newspaper: A court in Broward County, Florida, is starting to use an artificial intelligence (AI) program to help them decide if a defendant can be released on bail before trial. Early career judges are taking turns receiving advice from this AI program and another human judge, hired to serve as an advisor.*

We employed a factorial survey design [107] and showed participants eight vignettes that described a defendant from the ProPublica dataset, information about who the advisor was (i.e., an AI program or a human judge), which advice they gave, what the judge's final decision was, and whether the defendant committed a new crime within the next two years (i.e., re-offended). All vignettes stated that the judges' final decision *followed* the advice given, given the

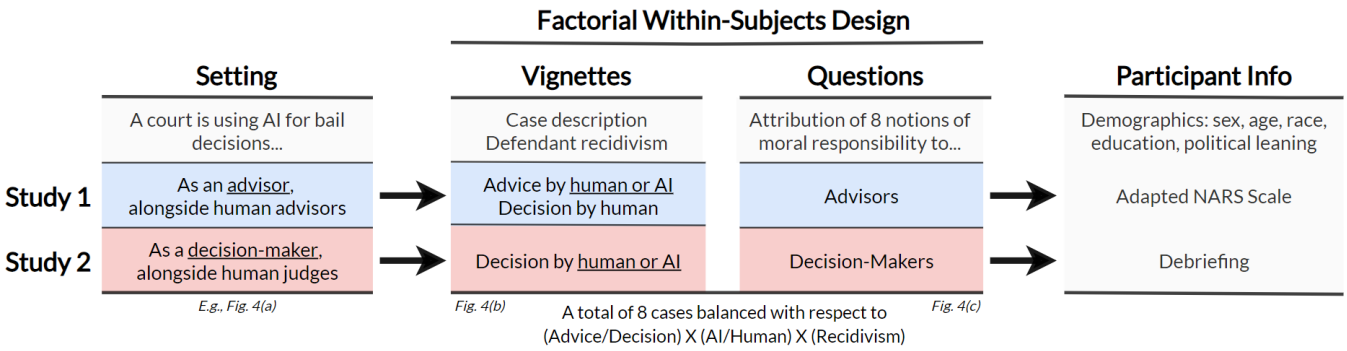


Figure 1: Survey Instrument. In Study 1, where AI advisors or human advisors assist human judges, survey respondents were asked to assign responsibility notions to AI and human advisors. In Study 2, where AI systems are decision-makers alongside human judges, survey respondents were asked to assign responsibility notions to AI and human decision-makers. Both studies employed a factorial within-subjects design that presented eight different vignettes to each respondent. Survey instruments are shown in Figure 4 of the Appendix.

ProPublica dataset does not provide this information. After reading the stimulus material, respondents were asked to indicate to what extent they agreed with a set of statements, presented in random order between participants, regarding the *advisor* on a 7-point Likert scale (-3 = Strongly Disagree, 3 = Strongly Agree).[2] These statements aimed to capture different notions of responsibility (see Table 2 in the Appendix for the complete list). Figure 1 illustrates the survey methodology. Participants were also asked two attention check questions in between vignettes.

Each participant in the study was exposed to a random subset of four cases with human advice and another four with AI advice. We ensured a balanced set was shown to each participant in terms of the advice (i.e., grant bail vs. deny bail) and recidivism. As a result, each respondent was shown one vignette of every possible combination of scenarios, encompassing eight (advice × recidivism × AI vs. human) variations. All vignettes were presented in random order to eliminate any order effect [44, 79].

Bail decisions aim to procure a balance between protecting future victims, e.g., prevent further offenses, and to impede any unnecessary burdens towards the defendant, e.g., by ensuring that their rights are protected [43]. The latter aspect of bail decisions is related to the assumption that one is innocent until proven otherwise beyond a reasonable doubt under criminal law [30]. To strike a balance between these two functions of bail decisions, we phrase statements addressing all notions of responsibility addressed in this work in two different forms: a human agent or an AI program could be held responsible for *i*) (not) protecting the rights of the defendant and *ii*) (not) preventing re-offense. Participants were randomly assigned to one of these treatment groups, and all statements followed the same phrasing style.

Towards the end of the survey, we asked demographic questions (presented in Table 1). We also gathered responses to a modified questionnaire of NARS (Negative Attitude towards Robot Scale) [92], whose subscale addressed "artificial intelligence programs" rather than "robots" to accommodate the COMPAS tool.

• Study 2: AI as Decision-Maker

Unlike Study 1, where a human decision-maker is advised by either a human or an AI advisor, Study 2 explores a setting that has yet to be implemented in the real-world. We imagine the case where an AI algorithm *makes* a bail decision by itself. The survey instrument and experimental design are identical to Study 1, except that in the introductory text, we told participants, *"The court is taking turns employing human judges and this AI program when making bailing decisions,"* and updated the phrasing of the questions to match this setting accordingly. In each vignette, participants were asked to what extent they agreed with the eight notions of responsibility regarding the *decision-maker*, i.e., the AI program or the human judge, using the same 7-point Likert scale from Study 1. Both studies had been approved by the Institutional Review Board (IRB) at the first author's institution.

• Pilot Study for Validation: Cognitive Interview

We validated our survey instruments through a series of cognitive interviews. Cognitive interviews are a standard survey methodology approach for improving the quality of questionnaires [83]. During the interviews, respondents accessed our web-based survey questionnaire and were interviewed by the authors while completing the survey. We utilized a verbal probing approach [110], in which we tested the respondents' interpretation of the survey questions, asked them to paraphrase the questions, and if they found the questions easy or difficult to understand and answer.

We interviewed six demographically diverse respondents. Three respondents were recruited through the online crowdsourcing platform Prolific [74], while the other three were our colleagues, who had prior experience designing and conducting human-subject studies. After each interview, we iteratively refined our survey instrument based on the respondent's feedback. We stopped gathering new responses once the feedback stopped leading to new insights. This process led to two significant changes in our survey instrument

[2] Questions related to responsibility-as-liability were shown in scenarios where *i*) the defendant re-offended and the phrasing style addressed the prevention of re-offenses, or *ii*) the defendants were denied bail and did not re-offend within two years while the statements focused on protecting their rights. The phrases tackling praise and blame were presented depending on the advice/decision and recidivism.

Demographic Attribute	Study 1	Study 2	Census
Total respondents	203	197	-
Passed attention checks	200	194	-
Women	41.5%	40.7%	51.0%
0-18 years old	-	-	21.7%
18-24 years old	37.5%	30.4%	10.8%
25-34 years old	30.5%	34.0%	13.7%
35-44 years old	17.0%	18.6%	12.6%
45-54 years old	8.0%	7.2%	13.4%
55-64 years old	6.0%	6.7%	12.9%
65+ years old	1.0%	2.6%	14.9%
Prefer not to respond	-	0.5%	-
African American	4.5%	7.2%	13%
Asian	17.0%	21.6%	6%
Caucasian	58.5%	54.6%	61%
Hispanic	10.0%	8.2%	18%
Other/Prefer not to respond	10.0%	8.4%	4%
Bachelor's Degree or above	48.0%	52.0%	30%
Liberal	59.0%	58.2%	33%†
Conservative	14.5%	18.0%	29%†
Moderate	23.5%	21.6%	34%†
Other/Prefer not to respond	3.0%	2.2%	4%†

Table 1: Respondents' demographics compared to the 2016 U.S. Census [100] and Pew data (marked with †) [75].

design. Firstly, we adapted the vignette presentation, which was initially adapted from previous work [33]. Our respondents unanimously stated that they found information about defendants easier to read, understand, and use when presented in a tabular format (shown in Figure 4 in the Appendix). Secondly, we rephrased some of the statements about the notions of responsibility we address in this work so that survey respondents' understanding of these concepts is similar to the definitions introduced above.

3.4 Participants and Recruitment

We conducted a power analysis to calculate the minimum sample size. A Wilcoxon-Mann-Whitney two-tailed test, with a 0.8 power to detect an effect size of 0.5 at the significance level of 0.05, requires 67 respondents per treatment group. Hence, we recruited 400 respondents through the Prolific crowdsourcing platform [74] to compensate for attention-check failures. We targeted US residents who have previously completed at least 100 tasks on Prolific, with an approval rate of 95% or above. Each participant was randomly assigned to one of the two studies.

The respondents' demographics are shown in Table 1. Prior studies of online crowdsourcing platforms have found that respondent samples tend to be younger, more educated, and consist of more women than the general US population [50]. Compared to the 2016 US census [100], our respondents are indeed younger and more highly educated. However, both of our studies' samples have a smaller ratio of women than the US population. Asian ethnicity is slightly over-represented in our samples. Compared to Pew Research data on the US population's political leaning [75], our samples are substantially more liberal.

The respondents were remunerated US$10.5 for taking part in the cognitive interviews and US$1.66 for completing the online surveys. The cognitive interviews lasted less than 30 minutes, while the online surveys took 10.36±5.43 minutes. Hence, all study participants were paid above the US minimum wage.

4 RESULTS

4.1 Responsibility in Bail Decisions

Figure 2 shows how people attributed each notion of responsibility to AI and human agents in Study 1 (on the advisor role) and Study 2 (on the decision-maker role).

First, responsibility-as-answerability (i.e., the bar in the middle) was the notion ascribed the highest to both human and AI advisors and decision-makers, followed by responsibility-as-obligation, as-task, as-authority, and as-power (i.e., the first four bars). On the other hand, liability and blame were the least attributed responsibility notion in bail decisions. Responsibility-as-cause and praise were the most neutral notions, and their mean attribution is close to zero (i.e., the baseline) across all treatments (see Figure 5 in the Appendix).

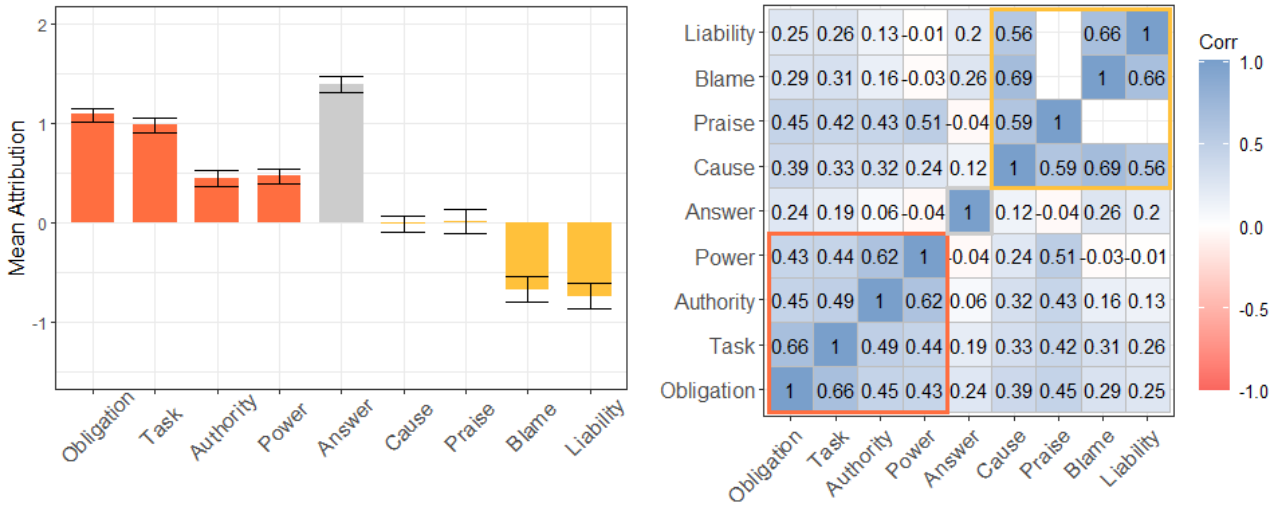
Second, Figure 2 shows two distinct sets of responsibility notions, where these clusters can be observed from the pairwise Spearman's correlation chart. A high correlation value indicates that those responsibility notions are perceived similarly by people. One group includes responsibility-as-task, as-authority, as-power, and as-obligation, all of which have positive mean values. The other group includes responsibility-as-cause, praise, blame, and liability. Responsibility-as-answerability belongs to neither of these groups.

Third, we can quantify variations across vignette conditions. Each vignette shown to participants varied in the advice given, bail decision, and recidivism, allowing us to compare across these factors. Our data show that vignettes that grant bail (as opposed to denying bail) led to a higher assignment of all responsibility notions, particularly causal responsibility and blame (see Figure 5 in the Appendix). A similar effect was found depending on defendant recidivism. For instance, the first four responsibility notions were ascribed to a more considerable degree if the defendant did not re-offend. In contrast, responsibility-as-cause, blame, and liability were attributed to a lesser extent if the defendant re-offended within two years. These trends corroborate the responsibility clusters discussed above.

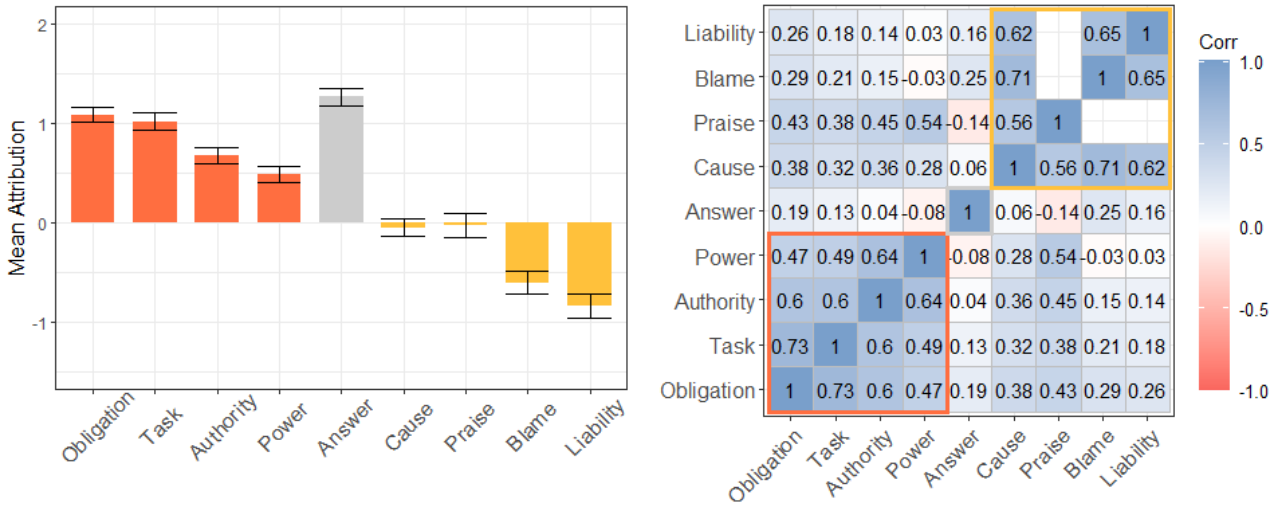
Finally, our study participants were also assigned to one of two different phrasing styles addressing some of the bailing decisions' objectives. Except for responsibility-as-answerability, addressing the violation or protection of a defendant's rights led to a marginally higher assignment of responsibility than the phrasing style focusing on preventing re-offenses.

4.2 Responsibility Assignment to AI and Humans

Our primary goal was to examine how people attribute responsibility to human and AI agents in high-stakes scenarios. To quantify the difference, we used a multivariate linear mixed model that included a random-effects term to control for each participant. This allowed us to account for repeated measures, i.e., explicitly model that each participant responded to questions on eight distinct defendants. We use the standard .05 level of significance. In all models, we use our adapted scale of pre-attitude towards AI systems as a



(a) Study 1: Aggregate results of AI or human advisors.



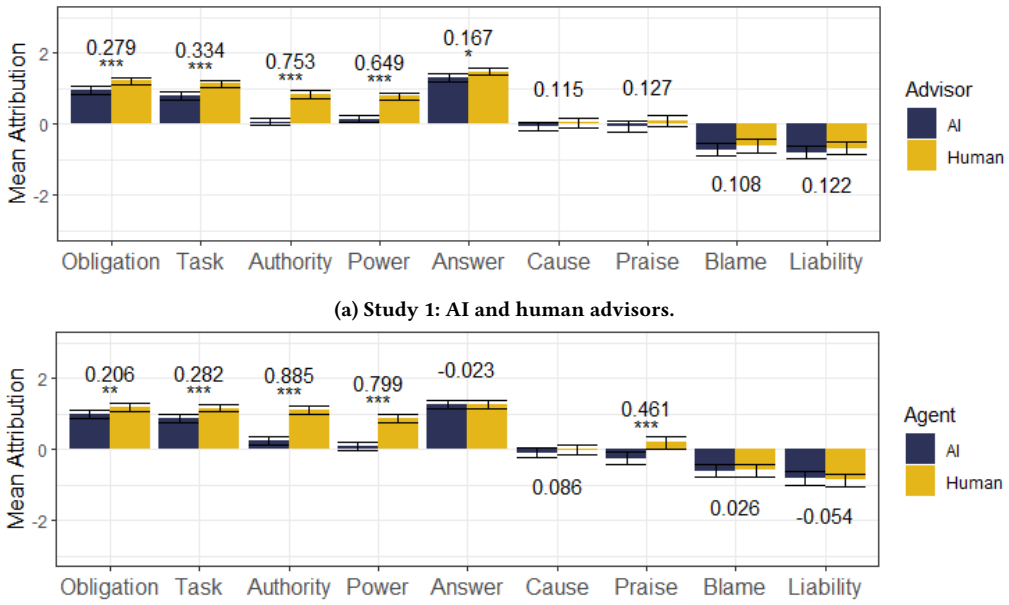
(b) Study 2: Aggregate results of AI or human decision-makers.

Figure 2: The overall attribution of responsibility to AI or human agents in bail decisions (left) and the correlation matrix across different responsibility notions (right). The y-axis indicates the degree to which participants attributed each notion of responsibility, based on a 7-pt Likert Scale (-3 = Strongly Disagree, 3 = Strongly Agree).

control variable. Figure 3 shows the results. The annotated numbers indicate the differences and significance levels between the two agents. We report the full regression coefficients in Table 3 in the Appendix.

Both Study 1 and Study 2 show consistent differences in responsibility attribution between agents, regardless of whether they informed a human judge (Study 1) or decided by themselves (Study 2). We note subtle differences in how people attribute responsibility to AI and humans. The first four responsibility concepts are correlated; the notions addressing tasks, supervisory roles, and the skills needed to assume them show a meaningful difference between agent types. The respondents attributed more of these notions of responsibility to humans than to AIs.

Responsibility-as-answerability exhibits a marginal difference with respect to the agent type that assisted human judges in bail decisions; however, the same trend was not observed in Study 2. Nevertheless, our results suggest that humans and AI are judged similarly responsible with respect to causality, blame, and liability for bail decisions. Moreover, human decision-makers are praised to a considerably larger degree than AI decision-makers, although the same effect was not observed for human and AI advisors.



(a) Study 1: AI and human advisors.

(b) Study 2: AI and human decision-makers.

Figure 3: Differences in responsibility attribution to AI programs and humans for bail decisions. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

5 DISCUSSION

5.1 The Relation Between Notions of Responsibility

So far, we have observed two clusters of responsibility concepts by their correlation. The first cluster is composed of responsibility-as-task, authority, power, and obligation — all of which were attributed to a greater degree to humans than AI systems (Δ>0.206, p<.001). The first three are descriptive and focus on one's tasks (i.e., task, authority) and the necessary skills for their completion (i.e., power). Furthermore, responsibility-as-obligation is related to responsibility-as-task in prescribing a specific goal to the agent; it differs from the latter, however, in setting a supervisory role towards the task, rather than specifying that one should be the one to complete it.

The second cluster includes causal responsibility, blame, praise, and liability — all of which were attributed to a similar degree to humans and AI. This finding is in line with previous work on blame assignment, highlighting the significance of causality in people's ascription of blame and punishment. Human subject studies suggest that blame attribution is a two-step process; it is initiated by a causal connection between an agent's action and its consequences and is followed by evaluating its mental states, i.e., intentions [25]. Malle et al. [66] have also proposed a theory of blame that is dependent on the causal connection between an agent and a norm-violating event. Our data similarly reveal such a relationship, even when controlling for the advice given, bail decision, or re-offense.

Concerning the phrasing styles, our experiment design addressed responsibility-as-liability as the duty to compensate those harmed by a wrongful action. However, previous work on the connection between liability (i.e., punishment) and causality focuses on the retributive aspect of punishment [25], often drawing a connection between punishment and blame. Therefore, we do not posit that people's ascription of liability is solely dependent on causality determinations. We hypothesize that the low assignment of liability is due to the current study's bail decision-making context. For instance, those wrongfully convicted do not receive any compensation for years spent in prison in at least 21 US states [88]. Hence, people might not believe that compensation is needed or deserved, or attribute this notion of responsibility to other entities, such as the court or the government, leading to a lower ascription of liability to the advisor or decision-maker.

Our findings indicate that participants who were presented with responsibility statements addressing the violation or protection of a defendant's rights (e.g., "It is the AI program's task to protect the rights of the defendant") were assigned higher responsibility levels across all notions. We posit that this effect results from the control that judges (humans and AIs) have over the consequences of their advice and decisions. Although a judge's decision can directly affect a defendant's rights depending on the appropriateness of one's jailing, preventing re-offenses is a complex task that encompasses diverse factors, such as policing and the defendant's decision to re-offend.

5.2 Humans Are More Responsible for Their Tasks Than AI Programs

Participants perceived human judges and advisors as more responsible for their tasks than their AI counterparts (see the leftmost bars in Figure 3). Humans are responsible for the tasks they are assigned, e.g., preventing re-offenses because they are in charge (i.e., authority) and have the skills necessary for completing them (i.e., power). These agents should either oversee (i.e., obligation) these

tasks or take the lead (i.e., task). On the other hand, AI systems are ascribed lower levels of all these responsibility notions.

The meanings of responsibility addressing the attribution of tasks and their requirements are descriptive in the sense that they should be addressed in the present tense [27], e.g., one *is* responsible for a task, or *is* in charge of it. Although descriptive and present-looking, these notions lead to the prescription of forward-looking responsibilities, such as an obligation. For instance, to be responsible for a specific task because one has the authority and necessary skills prescribes that one should see to it that the task is completed, i.e., an obligation is prescribed, through consequentialist, deontological, or virtue-based routes [102].

Participants attributed lower levels of authority and power to AI. This indicates that these systems are not thought to possess the necessary abilities to make decisions and advise such high-stakes decisions. Therefore, it is not deemed the AI program's responsibility to complete the assigned task or see it to be fulfilled.

5.3 The Need for Explanations

One of the prominent findings of this work is the need of interpretable AI systems. Although our participants assign a marginally lower level of responsibility-as-answerability for AI advisors vis-à-vis their human counterparts (Δ=0.167, p<.05), they believe they should justify their decisions to the same extent as human judges, particularly if they are to make the final bail decision (p>.05).

Moreover, our results suggest that an AI without a human-in-the-loop, i.e., AI judges in Study 2, could be held at the same level of scrutiny as human decision-makers for their decisions. This finding may imply that deploying black box AI in high-stakes scenarios, such as bail decision-making, will not be perceived well by the public. There exists empirical evidence that people might be averse to machines making moral decisions [10]. Previous work has not controlled for a system's interpretability, and therefore such trends might either *i*) be caused by the lack of explanations or *ii*) be aggravated if people become aware that AI systems cannot justify their moral decisions.

Judges should base their decisions on facts and be able to explain why they made such decisions. AI systems should also be capable of justifying their advice and decision-making process according to our results. This finding demonstrates the significance of these systems' interpretability. Scholars have discussed the risks posed by the opacity of existing AI algorithms. They argue that understanding how these systems come to their conclusions is necessary for both safe deployment and wide adoption [36]. Explainable AI (XAI) [46] is a field of computer science that has been given much attention in the community [38], and our results suggest that people agree with its importance.

Previous work has found that one's normative and epistemological values influence how explanations are comprehended [64]. Explanations involve both an explainer and explainee, meaning that conflicts might arise concerning how they are evaluated [69]. Therefore, we also posit that future work should delve deeper into what types of explanations the general public expects from AI systems. We highlight that those who are in charge of developing interpretable systems should not try to "nudge" recipients so they can be manipulated [63], e.g., for agency laundering [82].

5.4 AI and Human Agents Are Similarly Responsible for Consequences

The four rightmost bars in Figure 3 suggest that AI and human agents are ascribed similar levels of backward-notions of responsibility, namely blame, liability, praise, and causal responsibility.

5.4.1 The Relation Between Causality and Blame. A model that can explain our blameworthiness results is the Path Model of Blame, which proposes that blame is attributed through nested and sequential judgments of various aspects of the action and its agent [66]. After identifying a norm-violating event, the model states that one judges whether the agent is causally connected to the harmful outcome. If this causal evaluation is not successful, the model assigns little or no blame to the agent. Otherwise, the blamer evaluates the agent's intentionality. If the action is deemed intentional, the blamer evaluates the reasons behind it and ascribes blame accordingly. For unintentional actions, however, one evaluates whether the agent should have prevented the norm-violating event (i.e., had an obligation to prevent it) and could have done so (i.e., had the skills necessary), hence blaming the agent depending on the evaluation of these notions.

Our results from both studies show that AI and human agents are blamed to a similar degree. These findings agree with the Path Model of Blame, which proposes causality as the initial step for blame mitigation. The model proposes that one can mitigate blame by i) challenging one's causal connection to the wrongful action or ii) defending that it does not meet moral eligibility standards. We posit that the first excuse can explain why people blame human and AI advisors and decision-makers similarly. As their causal connection to the consequence is deemed alike, they are attributed to similar blame levels. Challenging one's causal effect in an outcome has also been discussed as a possible excuse to avoid blame by other scholars [101].

5.4.2 Praise in AI-Assisted Bail Decisions. The extent to which praise was assigned to human and AI agents varied depending on whether one was an advisor or a decision-maker. Even though Study 1 shows no difference between the two (p>.05), human decision-makers were more highly praised than AIs in Study 2 (Δ=0.461, p<.001). Previous work has proposed praise as a positive reinforcement [51] and a method through which one might convey information about its values and expectations to the praisee [28].

Regarding the difference between advisors and decision-makers, we posit that the differences between human agents are caused by the level of control the latter has over its decision outcomes. Although an advisor influences the final decision, the judge is the one who acts on it and, hence, deserves praise. Moreover, taking praise as positive reinforcement, praising the decision-maker over an advisor might have a bigger influence over future outcomes.

However, our results also indicate that AI decision-makers are not praised to the same level as human judges. Taking praise as a method of conveying social expectations and values, we highlight that people might perceive existing praising practices as inappropriate for AI. Similarly to the arguments against holding AI responsible per se, focusing on the fact that they do not have mental states required for existing responsibility practices [89, 96], praising an AI might lose its meaning if done as if it were towards humans.

The same argument could also be applied to the practice of blame [26]. If the general public believes praising an AI system does not make sense, people might perceive blameworthiness similarly, contradicting our results. However, studies have shown a public impulse to blame, driven by the desire to express social values and expectations [18]. Psychological evidence further suggests that humans are innate retributivists [17]. Likewise, HCI research has found that people attribute blame to robotic agents upon harm, particularly if they are described to be autonomous and serve the main cause of harm [37, 52, 67]. Hence, there is no contradiction in people attributing blame to AI systems for harms, although they should not be praised for opposing consequences.

5.4.3 Liability as Compensation. Our findings indicating that AI and human agents should be held liable to a similar level goes against previous work, which has found that people attribute punishment to AI systems to a lesser degree than their human counterparts [62]. Punishment fulfills many societal goals, such as making victims whole, the satisfaction of retributive feelings, and offenders' reform. In the current study, we address one of these functions and phrase liability as the responsibility to compensate those harmed (i.e., make victims whole). Therefore, our results do not directly contradict earlier findings that had addressed punishment in its wide definition.

The results from our initial exploratory analysis in Section 4.1 show that trends found between causality and blame attributions across different phrasing styles do not directly transfer to liability judgments. Hence, we do not posit that similar causality judgments can explain the similar attribution of liability to AI and humans as in the case of blame. Still, we instead hypothesize that it results from two different factors based on our phrasing styles.

Regarding the statements addressing the prevention of re-offenses, we posit that the lower attribution of liability to both agents is caused by a variation of the "problem of many hands." [102] Preventing defendants from re-offending does not rely solely on a judge's decision but encompasses many other factors as discussed above. Therefore, liability is distributed across various entities, such as the government and the court per se. Addressing the statements focusing on protecting defendants' rights, we hypothesize that people do not expect defendants to be compensated if their rights are violated. As examined above, much of the US legislature does not compensate those who have been unjustly incarcerated [88]. The respondents did not believe those harmed should, or even could, be made whole for the violation of their rights, and hence, both AI and human agents are attributed low and similar levels of liability.

6 IMPLICATIONS

Our findings indicate that people believe humans are, and should be, responsible for the assigned tasks, regardless of whether they are advisors or decision-makers. Our respondents perceive humans as having the skills necessary to complete these tasks, being in charge of them, and being able to ensure that they are completed. The responsibility notions that were attributed to human agents to a greater extent than to AIs are present- and forward-looking in the sense that they are descriptive, i.e., by stating a fact, and prescribe obligations. It is important to note that users of AI systems are also responsible in a backward-looking fashion such that they should also be held responsible for the outcomes of their advice and decisions. Therefore, our findings agree with scholars who propose that users (and designers) should take responsibility for their automated systems' actions and consequences [20, 72].

Nonetheless, our study shows that AIs could also be held responsible for their actions. Taking morality as a human-made construct [93], it may be inevitable to hold AI systems responsible alongside their users and designers so that this formulation is kept intact. Viewing responsibility concepts as social constructs that aim to achieve specific social goals, attributing backward-looking notions of responsibility to AI systems might emphasize these goals [91]. Our study indicates these practices might not need to focus on compensating those harmed by these systems given the low attribution of liability to all agents.[3] We instead hypothesize that people might desire to hold these entities responsible for retributive motives, such as satisfying their needs for revenge [71] and bridging the retribution gap [26], as a result of human nature [25]. It is important to note that AI systems might not be appropriate subjects of (retributive) blame [26, 89], i.e., scholars argue that blaming automated agents would be wrong and unsuccessful. Future research can address which functions of responsibility attribution would satisfy this public attribution of backward-looking responsibilities to AI systems. Future studies can also address scenarios in which blame could be attributed to a higher degree, e.g., those with life-or-death consequences, such as self-driving vehicles and AI medical advisors.

A common concern raised by scholarly work is that blaming or punishing an AI system might lead to social disruptions. From a legal perspective, attributing responsibility to these systems might obfuscate designers and users' roles, creating human liability shields [16], i.e., stakeholders might use these automated systems as a form of protecting themselves from deserved punishment. Another possible issue is "agency laundering," in which the systems' designer distances itself from morally suspect actions, regardless of intentionality, by blaming the algorithm, machine, or system [82]. This form of blame-shifting has been observed, for example, when Facebook called out its algorithm for autonomously creating anti-semitic categories in its advertisement platform [1, 97]. We highlight that any responsibility practice towards AI systems should not blur the responsibility prescribed and deserved by their designers and users. Our findings suggest that autonomous algorithms alone should not be held responsible by themselves, but rather *alongside* other stakeholders, so these concerns are not realized.

7 CONCLUDING REMARKS

This paper discussed the responsibility gap posed by the deployment of autonomous AI systems [68] and conducted a survey study to understand how differently people attribute responsibility to AI and humans. As a case study, we adapted vignettes from real-life algorithm-assisted bail decisions and employed a within-subjects experimental design to obtain public perceptions on various notions of moral responsibility. We conducted two studies; the former illustrated a realistic scenario in which AI advises human judges,

[3] This finding does not imply that those harmed should not be compensated, but rather that respondents do not attribute this responsibility to AI systems per se. Some scholars propose that other stakeholders should take this responsibility [19], mainly because automated agents are not capable of doing so [16].

and the latter described a fictional circumstance where AI is the decision-maker itself.

The current study focused on AI systems currently being used to advise bailing decisions, which is an important yet specific application of these algorithms. Therefore, our results might not be generalizable to all possible environments. For instance, some of our results partly conflict with previous work addressing self-driving vehicles [7] and medical systems [62]. Studies such as ours should be expanded to diverse AI applications, where they are used both in-the-loop (as in Study 1) and autonomously (as in Study 2). People have different opinions regarding how (and where) these systems should be deployed in relation to how autonomous they should be [65], which should affect how they ascribe responsibility for their actions.

Study 1 was designed so that the judge's decision always followed the advice given to reduce complexity in the vignette design. However, future studies on similar topics should also consider scenarios in which AI systems and their human supervisors disagree. For instance, if a human judge chooses to disagree with advice, some of the advisor's responsibilities might be shifted towards the decision-maker regardless of the advisor's nature. In our case study, human-AI collaborations are such that there exists an AI-in-the-loop; future work should address other collaboration variations, such as human-in-the-loop, i.e., humans assisting machines.

The current research considered eight notions of responsibility from related work. We recognize that other meanings of responsibility could be further considered, such as virtue-based notions where one might call an entity responsible in that it prescribes an evaluation of one's traits and dispositions [87, 94]. These notions have been widely agreed upon as incompatible with AI systems due to their lack of metaphysical attributes [20, 89, 96]. Nevertheless, our research has found key clusters of responsibility notions concerning AI and human agents, opening further research directions.

Our exploratory analysis identified two clusters of responsibility notions. One cluster encompasses meanings related to the attribution of tasks and obligations (i.e., responsibility-as-task, as-obligation), their necessary skills (i.e., responsibility-as-power), and the ascription of authority (i.e., responsibility-as-authority). The other cluster includes meanings related to causal determinations (i.e., responsibility-as-cause) and backward-looking responsibility notions (i.e., blame, praise, and liability).

As our results demonstrate, people may hold AI to a similar level of moral scrutiny as humans for their actions and harms. Our respondents indicate that they expect decision-makers and advisors to justify their bailing decisions regardless of their nature. Our findings highlight the importance of interpretable and explainable algorithms, particularly in high-stakes scenarios, such as our case-study. Finally, this study also showed that people judge AI and humans differently with respect to certain notions of responsibility, particularly those addressing present- and forward-looking meanings, such as responsibility-as-task and as-obligation. However, we have also found that people attribute similar levels of causal responsibility, blame, and liability to AI and human advisors and decision-makers for bail decisions.

ACKNOWLEDGMENTS

This work was supported by the Institute for Basic Science (IBS-R029-C2).

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2017. Facebook Enabled Advertisers to Reach 'Jew Haters'. https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters.
[2] Julia Angwin, Madeleine Varner, and Ariana Tobin. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[3] Peter M Asaro. 2006. What should we want from a robot ethic? *The International Review of Information Ethics* 6 (2006), 9–16.
[4] Peter M Asaro. 2011. A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics. *Robot ethics: The ethical and social implications of robotics* (2011), 169.
[5] Peter M Asaro. 2016. The liability problem for autonomous artificial agents. In *2016 AAAI Spring Symposium Series*.
[6] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59–64.
[7] Edmond Awad, Sydney Levine, Max Kleiman-Weiner, Sohan Dsouza, Joshua B Tenenbaum, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2020. Drivers are blamed more than their automated cars when both make mistakes. *Nature Human Behaviour* 4, 2 (2020), 134–143.
[8] Susanne Beck. 2016. The problem of ascribing legal responsibility in the case of robotics. *AI & Society* 31, 4 (2016), 473–481.
[9] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* (2018), 0049124118782533.
[10] Yochanan E Bigman and Kurt Gray. 2018. People are averse to machines making moral decisions. *Cognition* 181 (2018), 21–34.
[11] Gunnar Björnsson and Karl Persson. 2012. The explanatory component of moral responsibility. *Noûs* 46, 2 (2012), 326–354.
[12] J Bonnefon, Azim Shariff, and Iyad Rahwan. 2020. *The moral psychology of AI and the ethical opt-out problem.* Oxford University Press, Oxford, UK.
[13] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352, 6293 (2016), 1573–1576.
[14] Marcus Alphons Petrus Bovens, Warwick Ford, Mark Bovens, et al. 1998. *The quest for responsibility: Accountability and citizenship in complex organisations.* Cambridge university press.
[15] Bartosz Brożek and Marek Jakubiec. 2017. On the legal responsibility of autonomous machines. *Artificial Intelligence and Law* 25, 3 (2017), 293–304.
[16] Joanna J Bryson, Mihailis E Diamantis, and Thomas D Grant. 2017. Of, for, and by the people: the legal lacuna of synthetic persons. *Artificial Intelligence and Law* 25, 3 (2017), 273–291.
[17] Kevin M Carlsmith. 2008. On justifying punishment: The discrepancy between words and actions. *Social Justice Research* 21, 2 (2008), 119–137.
[18] Kevin M Carlsmith, John M Darley, and Paul H Robinson. 2002. Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of personality and social psychology* 83, 2 (2002), 284.
[19] Paulius Čerka, Jurgita Grigienė, and Gintarė Sirbikytė. 2015. Liability for damages caused by artificial intelligence. *Computer Law & Security Review* 31, 3 (2015), 376–389.
[20] Marc Champagne and Ryan Tonkens. 2015. Bridging the responsibility gap in automated warfare. *Philosophy & Technology* 28, 1 (2015), 125–137.
[21] Eugene Chislenko. 2019. Scanlon's Theories of Blame. *The Journal of Value Inquiry* (2019), 1–16.
[22] Cory J Clark, Eric Evan Chen, and Peter H Ditto. 2015. Moral coherence processes: Constructing culpability and consequences. *Current Opinion in Psychology* 6 (2015), 123–128.
[23] Mark Coeckelbergh. 2009. Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents. *AI & Society* 24, 2 (2009), 181–189.
[24] Mark Coeckelbergh. 2019. Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and engineering ethics* (2019), 1–18.
[25] Fiery Cushman. 2008. Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition* 108, 2 (2008), 353–380.
[26] John Danaher. 2016. Robots, law and the retribution gap. *Ethics and Information Technology* 18, 4 (2016), 299–309.
[27] Michael Davis. 2012. "Ain't no one here but us social forces": Constructing the professional responsibility of engineers. *Science and Engineering Ethics* 18, 1 (2012), 13–34.

[28] Catherine R Delin and Roy F Baumeister. 1994. Praise: More than just social reinforcement. *Journal for the theory of social behaviour* 24, 3 (1994), 219–241.

[29] Mady Delvaux. 2017. Report with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103 (INL)). *European Parliament Committee on Legal Affairs* (2017).

[30] Mandeep K Dhami, Samantha Lundrigan, and Katrin Mueller-Johnson. 2015. Instructions on reasonable doubt: Defining the standard of proof and the juror's task. *Psychology, Public Policy, and Law* 21, 2 (2015), 169.

[31] Virginia Dignum. 2017. Responsible artificial intelligence: designing AI for human values. (2017).

[32] Gordana Dodig-Crnkovic and Daniel Persson. 2008. Sharing moral responsibility with robots: A pragmatic approach. *Frontiers in Artificial Intelligence And Applications* 173 (2008), 165.

[33] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1 (2018).

[34] Robin Antony Duff. 2007. *Answering for crime: Responsibility and liability in the criminal law*. Bloomsbury Publishing.

[35] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (2017), 115–118.

[36] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, et al. 2018. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines* 28, 4 (2018), 689–707.

[37] Caleb Furlough, Thomas Stokes, and Douglas J Gillan. 2019. Attributing Blame to Robots: I. The Influence of Robot Autonomy. *Human Factors* (2019), 0018720819880641.

[38] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 80–89.

[39] Sabine Gless, Emily Silverman, and Thomas Weigend. 2016. If Robots cause harm, Who is to blame? Self-driving Cars and Criminal Liability. *New Criminal Law Review* 19, 3 (2016), 412–436.

[40] Kurt Gray, Chelsea Schein, and Adrian F Ward. 2014. The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General* 143, 4 (2014), 1600.

[41] Nina Grgic-Hlaca, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference*. 903–912.

[42] Nina Grgić-Hlača, Adrian Weller, and Elissa M Redmiles. 2020. Dimensions of Diversity in Human Perceptions of Algorithmic Fairness. (2020). arXiv:2005.00808

[43] Nina Grgić-Hlača, Christoph Engel, and Krishna P. Gummadi. 2019. Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 178 (Nov. 2019), 25 pages. https://doi.org/10.1145/3359280

[44] Robert M Groves, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau. 2011. *Survey Methodology*. Vol. 561. John Wiley & Sons.

[45] David J Gunkel. 2017. Mind the gap: responsible robotics and the problem of responsibility. *Ethics and Information Technology* (2017), 1–14.

[46] David Gunning. 2017. Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2 (2017), 2.

[47] Raul Hakli and Pekka Mäkelä. 2016. Robots, Autonomy, and Responsibility. (2016).

[48] F Allan Hanson. 2009. Beyond the skin bag: On the moral responsibility of extended agencies. *Ethics and Information Technology* 11, 1 (2009), 91–99.

[49] Zan Huang, Hsinchun Chen, Chia-Jung Hsu, Wun-Hwa Chen, and Soushan Wu. 2004. Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems* 37, 4 (2004), 543–558.

[50] Panagiotis G Ipeirotis. 2010. Demographics of mechanical turk. (2010).

[51] Alan E Kazdin. 1978. *History of behavior modification: Experimental foundations of contemporary research*. University Park Press.

[52] Taemie Kim and Pamela Hinds. 2006. Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 80–85.

[53] Sebastian Köhler, Neil Roughley, and Hanno Sauer. 2017. Technology, responsibility gaps and the robustness of our everyday conceptual scheme. *Moral Agency and the Politics of Responsibility* (2017).

[54] Bert-Jaap Koops, Mireille Hildebrandt, and David-Olivier Jaquet-Chiffelle. 2010. Bridging the accountability gap: Rights for new entities in the information society. *Minn. JL Sci. & Tech.* 11 (2010), 497.

[55] Francesca Lagioia and Giovanni Sartor. 2019. AI Systems Under Criminal Law: a Legal Analysis and a Regulatory Perspective. *Philosophy & Technology* (2019), 1–33.

[56] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684.

[57] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[58] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. 2019. WeBuildAI: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–35.

[59] Jamy Li, Xuan Zhao, Mu-Jung Cho, Wendy Ju, and Bertram F Malle. 2016. *From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with self-driving cars*. Technical Report. SAE Technical Paper.

[60] Dafni Lima. 2017. Could AI Agents Be Held Criminally Liable: Artificial Intelligence and the Challenges for Criminal Law. *SCL Rev.* 69 (2017), 677.

[61] Gabriel Lima and Meeyoung Cha. 2020. Responsible AI and Its Stakeholders. (2020). arXiv:2004.11434

[62] Gabriel Lima, Chihyung Jeon, Meeyoung Cha, and Kyungsin Park. 2020. Will Punishing Robots Become Imperative in the Future?. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8. https://doi.org/10.1145/3334480.3383006

[63] Zachary C Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (2018), 31–57.

[64] Tania Lombrozo. 2009. Explanation and categorization: How "why?" informs "what?". *Cognition* 110, 2 (2009), 248–253.

[65] Brian Lubars and Chenhao Tan. 2019. Ask not what AI can do, but what AI should do: Towards a framework of task delegability. In *Advances in Neural Information Processing Systems*. 57–67.

[66] Bertram F Malle, Steve Guglielmo, and Andrew E Monroe. 2014. A theory of blame. *Psychological Inquiry* 25, 2 (2014), 147–186.

[67] Bertram F. Malle, Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. 2015. Sacrifice One For the Good of Many? People Apply Different Moral Norms to Human and Robot Agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '15)*. https://doi.org/10.1145/2696454.2696458

[68] Andreas Matthias. 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology* 6, 3 (2004), 175–183.

[69] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* (2019), 1–7.

[70] Andrew E Monroe and Bertram F Malle. 2017. Two paths to blame: Intentionality directs moral information processing along two distinct tracks. *Journal of Experimental Psychology: General* 146, 1 (2017), 123.

[71] Christina Mulligan. 2017. Revenge against robots. *SCL Rev.* 69 (2017), 579.

[72] Sven Nyholm. 2018. Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and Engineering Ethics* 24, 4 (2018), 1201–1219.

[73] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[74] Stefan Palan and Christian Schitter. 2018. Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27.

[75] Pew Research Center. 2016. 2016 Party Identification Detailed Tables. http://www.people-press.org/2016/09/13/2016-party-identification-detailed-tables/

[76] Emma Pierson. 2017. Gender differences in beliefs about algorithmic fairness. (2017). arXiv:1712.09124

[77] William Lloyd Prosser et al. 1941. *Handbook of the Law of Torts*. Vol. 4. West Publishing.

[78] Iyad Rahwan. 2018. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology* 20, 1 (2018), 5–14.

[79] Elissa M Redmiles, Yasemin Acar, Sascha Fahl, and Michelle L Mazurek. 2017. *A Summary of Survey Methodology Best Practices for Security and Privacy Researchers*. Technical Report.

[80] Neil M Richards and William D Smart. 2016. How should the law think about robots? In *Robot Law*. Edward Elgar Publishing.

[81] Lionel P Robert, Casey Pierce, Liz Marquis, Sangmi Kim, and Rasha Alahmad. 2020. Designing fair AI for managing employees in organizations: a review, critique, and design agenda. *Human–Computer Interaction* (2020), 1–31.

[82] Alan Rubel, Clinton Castro, and Adam Pham. 2019. Agency Laundering and Information Technologies. *Ethical Theory and Moral Practice* 22, 4 (2019), 1017–1041.

[83] Katherine Ryan, Nora Gannon-Slater, and Michael J Culbertson. 2012. Improving survey methods with cognitive interviews in small-and medium-scale evaluations. *American Journal of Evaluation* 33, 3 (2012), 414–430.

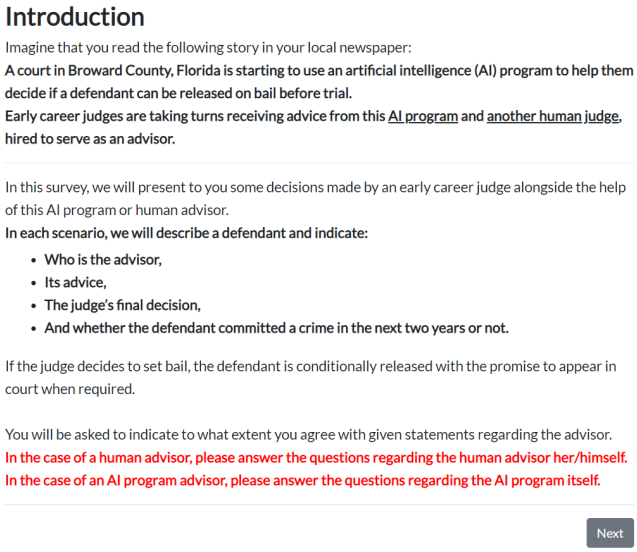
[84] Nripsuta Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David Parkes, and Yang Liu. 2019. How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness. *AIES* (2019).

[85] Thomas Scanlon. 2000. *What we owe to each other.* Belknap Press.
[86] Thomas M Scanlon. 2008. Moral dimensions: Meaning, permissibility, and blame. *Cambridge: Harvard* (2008).
[87] David Shoemaker. 2011. Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics* 121, 3 (2011), 602–632.
[88] Stephanie Slifer. 2014. *How the wrongfully convicted are compensated for years lost.* Available at https://www.cbsnews.com/news/how-the-wrongfully-convicted-are-compensated/. Last accessed September 4 2020.
[89] Robert Sparrow. 2007. Killer robots. *Journal of Applied Philosophy* 24, 1 (2007), 62–77.
[90] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. (2019). arXiv:1902.04783
[91] Bernd Carsten Stahl. 2006. Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency. *Ethics and Information Technology* 8, 4 (2006), 205–213.
[92] Dag Sverre Syrdal, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L Walters. 2009. The negative attitudes towards robots scale and reactions to robot behaviour in a live human-robot interaction study. *Adaptive and Emergent Behaviour and Complex Systems* (2009).
[93] Andreas Theodorou. [n.d.]. Why Artificial Intelligence is a Matter of Design. *Reflections in Philosophy, Theology, and the Social Sciences* ([n. d.]), 105.
[94] Daniel W Tigard. 2020. Responsible AI and moral responsibility: a common appreciation. *AI and Ethics* (2020), 1–5.
[95] Daniel W Tigard. 2020. There Is No Techno-Responsibility Gap. *Philosophy & Technology* (2020), 1–19.
[96] Steve Torrance. 2008. Ethics and consciousness in artificial agents. *Ai & Society* 22, 4 (2008), 495–521.
[97] Andreas Tsamados, Nikita Aggarwal, Josh Cowls, Jessica Morley, Huw Roberts, Mariarosaria Taddeo, and Luciano Floridi. 2020. The Ethics of Algorithms: Key Problems and Solutions. *Available at SSRN 3662302* (2020).
[98] Jacob Turner. 2018. *Robot rules: Regulating artificial intelligence.* Springer.
[99] Mathias Twardawski, Karen TY Tang, and Benjamin E Hilbig. 2020. Is It All About Retribution? The Flexibility of Punishment Goals. *Social Justice Research* (2020), 1–24.
[100] U.S. Census Bureau. 2016. American Community Survey 5-Year Estimates.
[101] Ibo Van de Poel. 2011. The relation between forward-looking and backward-looking responsibility. In *Moral Responsibility.* Springer, 37–52.
[102] Ibo Van de Poel. 2015. Moral responsibility. In *Moral responsibility and the problem of many hands.* Routledge, 24–61.
[103] Robert van den Hoven van Genderen. 2018. Do we need new legal personhood in the age of robots and AI? In *Robotics, AI and the Future of Law.* Springer, 15–55.
[104] Nicole A Vincent. 2011. A structured taxonomy of responsibility concepts. In *Moral responsibility.* Springer, 15–35.
[105] Laura Wächter and Felix Lindner. 2018. An explorative comparison of blame attributions to companion robots across various moral dilemmas. In *Proceedings of the 6th International Conference on Human-Agent Interaction.* 269–276.
[106] R Jay Wallace. 1994. *Responsibility and the moral sentiments.* Harvard University Press.
[107] Lisa Wallander. 2009. 25 years of factorial surveys in sociology: A review. *Social Science Research* 38, 3 (2009), 505–520.
[108] Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–14.
[109] Michael Wenzel and Ines Thielmann. 2006. Why we punish in the name of justice: Just desert versus value restoration and the role of social identity. *Social Justice Research* 19, 4 (2006), 450–470.
[110] Gordon B Willis. 2004. *Cognitive interviewing: A tool for improving questionnaire design.* Sage Publications.
[111] Chen Zhu, Hengshu Zhu, Hui Xiong, Chao Ma, Fang Xie, Pengliang Ding, and Pan Li. 2018. Person-job fit: Adapting the right talent for the right job with joint representation learning. *ACM Transactions on Management Information Systems (TMIS)* 9, 3 (2018), 1–17.

A APPENDIX

Notion	Phrasing	Statements
Responsibility-as-task	Further Offense	It is the (agent)'s task to prevent further offenses.
	Rights	It is the (agent)'s task to protect the rights of the defendant.
Responsibility-as-authority	Further Offense	The (agent) has the authority to prevent further offenses.
	Rights	The (agent) has the authority to protect the rights of the defendant.
Responsibility-as-power	Further Offense	The (agent) has the skills needed to prevent further offenses.
	Rights	The (agent) has the skills needed to protect the rights of the defendant.
Responsibility-as-obligation	Further Offense	The (agent) should ensure that no further offense is committed.
	Rights	The (agent) should ensure that the rights of the defendant are protected.
Responsibility-as-answerability	Further Offense	The (agent) should justify their advice/decision.
	Rights	The (agent) should justify their advice/decision.
Responsibility-as-cause	Further Offense	The (agent)'s decision led to the occurrence/prevention of the reoffense.
	Rights	The (agent)'s decision led to the violation/protection of the rights of the defendant.
Responsibility-as-blame/praise	Further Offense	The (agent) should be blamed/praised for the failure to prevent/prevention of the reoffense.
	Rights	The (agent) should be blamed/praised for the violation/protection of the rights of the defendant.
Responsibility-as-liability	Further Offense	The (agent) should compensate those harmed by the reoffense.
	Rights	The (agent) should compensate the defendant for violating their rights.

Table 2: Statements addressing all responsibility notions presented to participants in Study 1 and Study 2. (Agent) is either "AI program," "human advisor," or "human judge" depending on the agent and the study. The statements addressing responsibility-as-liability were shown if *i*) the defendant re-offended and the phrasing style addressed the prevention of re-offenses, or *ii*) the defendants were denied bail and did not re-offend within two years while the statements focused on the protection of their rights. The phrases tackling praise and blame were presented depending on the advice/decision and recidivism. The phrasing column indicates how statements were phrased depending on which function of the bail decision they stressed: preventing further offenses (Further Offense) or protecting the defendant's rights (Rights).



(a) Study introduction presenting the scenario where AI systems are being used for bail decisions.

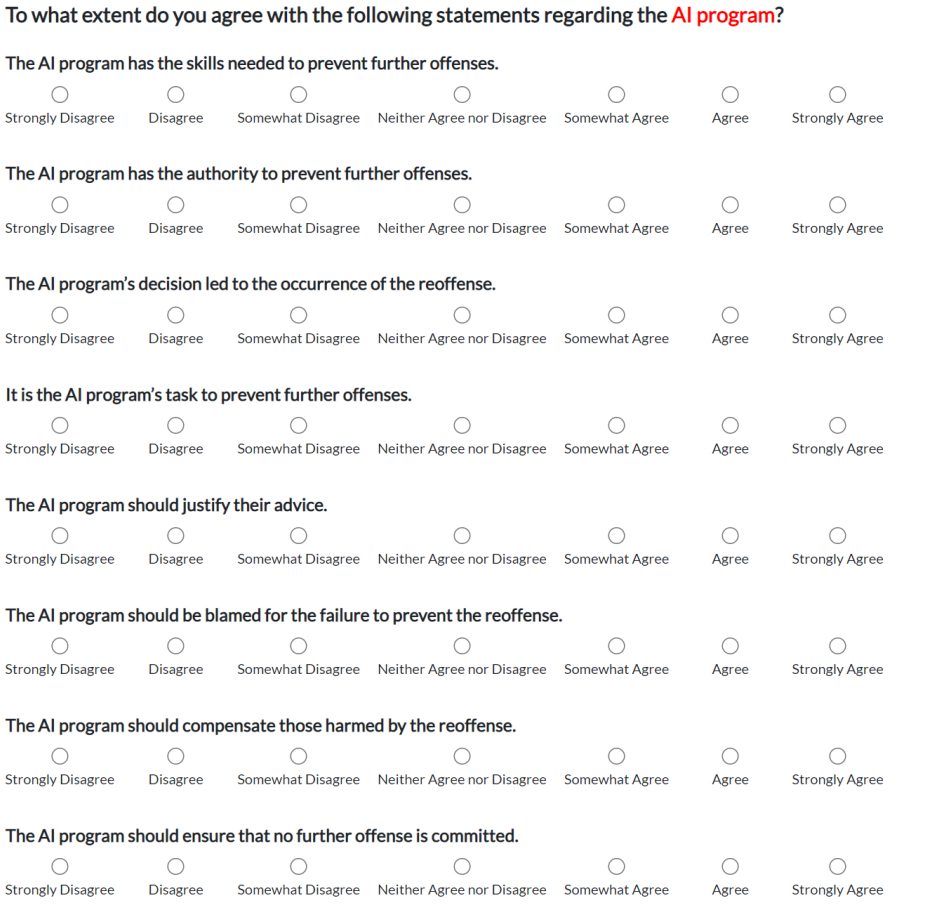
Sex	Male
Age	27 years old
Crime	Criminal Damage of less than $1000
Classification of Crime	Misdemeanor
Number of Prior Crimes	5
Number of Juvenile Felonies	0
Number of Juvenile Misdemeanors	0

AI program's advice: grant bail.

Human judge's decision: grant bail.

Defendant: reoffended during bail.

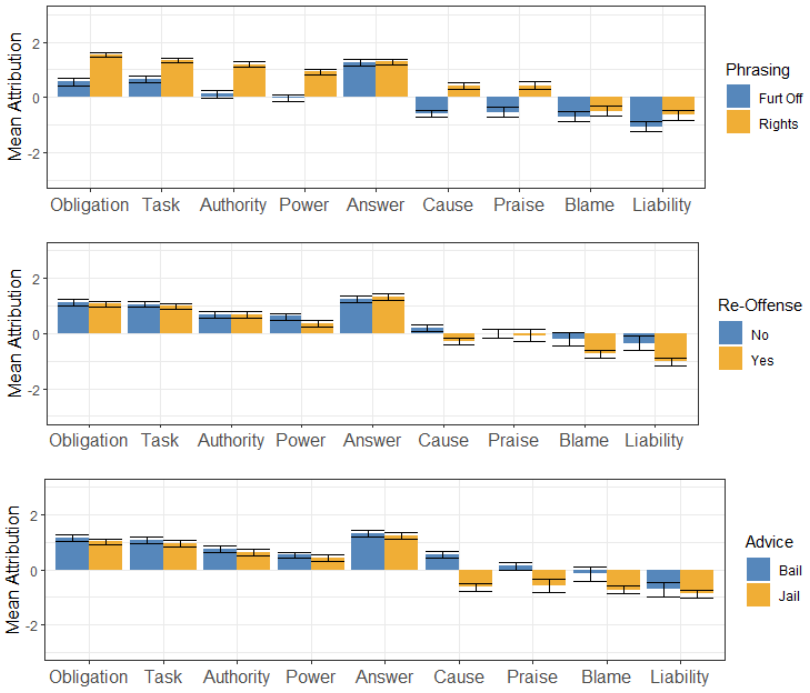
(b) Vignette presented to survey participants introducing a defendant, whether they have re-offended, and the stakeholders' decisions and advices.



(c) Attribution of the eight notions of moral responsibility to the advisor in Study 1 (or decision-maker in Study 2).

Figure 4: Example screenshots of the survey instrument used for Study 1. The study is available at https://thegcamilo.github.io/responsibility-compas/.

(a) Study 1: AI and human advisors.



(b) Study 2: AI and human decision-makers.

Figure 5: Attribution of responsibility for bail decisions depending on how the statements were phrased, recidivism, and advice/decision.

	(1) Decision-Maker	(2) Advisor
answer		
agent_human	-0.0232	0.168*
advice_jail	-0.0722	-0.0800
defendant_reoffended	0.0696	0.130
phrasing_rights	0.0308	-0.102
control	0.0103	-0.119
intercept	1.254***	1.501***
authority		
agent_human	0.885***	0.753***
advice_jail	-0.122*	-0.193**
defendant_reoffended	-0.00129	-0.233***
phrasing_rights	1.052***	0.702***
control	-0.150	-0.0790
intercept	-0.0410	0.0360
blame		
agent_human	0.0258	0.107
advice_jail	-1.128***	-0.878***
defendant_reoffended	-0.572***	-0.456**
phrasing_rights	0.500*	0.190
control	0.239*	-0.0312
intercept	0.0484	0.226
cause		
agent_human	0.0863	0.115
advice_jail	-1.179***	-1.018***
defendant_reoffended	-0.460***	-0.550***
phrasing_rights	1.020***	0.665***
control	0.0278	-0.0686
intercept	0.136	0.468**
liability		
agent_human	-0.0541	0.122
advice_jail	-0.722***	-0.541**
defendant_reoffended	-0.553***	-0.466**
phrasing_rights	0.530*	0.0999
control	0.232*	0.0355
intercept	-0.481	-0.145
obligation		
agent_human	0.206***	0.279***
advice_jail	-0.144**	-0.124*
defendant_reoffended	-0.0438	-0.216***
phrasing_rights	0.987***	0.815***
control	-0.0614	-0.0786
intercept	0.642***	0.814***
power		
agent_human	0.799***	0.649***
advice_jail	-0.101	-0.126
defendant_reoffended	-0.271***	-0.366***
phrasing_rights	0.939***	0.572***
control	-0.0759	-0.118
intercept	-0.119	0.263
praise		
agent_human	0.461***	0.128
advice_jail	-0.0611	0.0969
defendant_reoffended	-0.990***	-0.941***
phrasing_rights	1.435***	0.950***
control	0.0277	-0.120
intercept	-0.784***	-0.148
task		
agent_human	0.282***	0.334***
advice_jail	-0.117*	-0.111
defendant_reoffended	-0.0683	-0.164**
phrasing_rights	0.686***	0.345**
control	-0.0417	0.0128
intercept	0.667***	0.760***

Table 3: Coefficients from the multivariate mixed effects model presented in Section 4.2. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.