指责人类与机器：是什么塑造了人们对算法伤害的反应

Gabriel Lima
gabriel.lima@kaist.ac.kr
KAIST计算学院& IBS数据科学组
大韩民国

Nina Grgić-Hlača
nghlaca@mpi-sws.org 马克斯·普朗克软件系统研究所 & 马克斯·普朗克集体物品研究所 德国

Meeyoung Cha
mcha@ibs.re.kr 韩国基础科学研究院数据科学组及韩国科学技术院计算学院 大韩民国

arXiv:2304.02176v1 [cs.CY] 5 Apr 2023

摘要

Artificial intelligence (AI) systems can cause harm to people. This research examines how individuals react to such harm through the lens of blame. Building upon research suggesting that people blame AI systems, we investigated how several factors influence people's reactive attitudes towards machines, designers, and users. The results of three studies (N = 1,153) indicate differences in how blame is attributed to these actors. Whether AI systems were explainable did not impact blame directed at them, their developers, and their users. Considerations about fairness and harmfulness increased blame towards designers and users but had little to no effect on judgments of AI systems. Instead, what determined people's reactive attitudes towards machines was whether people thought blaming them would be a suitable response to algorithmic harm. We discuss implications, such as how future decisions about including AI systems in the social and moral spheres will shape laypeople's reactions to AI-caused harm.

CCS 概念

• **以人为中心的计算** → *HCI* 中的实证研究； • **应用计算** → **心理学**。

关键词

人工智能、责备、责任、可解释性、伤害、决策、歧视、算法、算法决策

ACM参考文献格式：

加布里埃尔·利马，以及 MeeyoungCha。2023。《责怪人类与机器：是什么塑造了人们对算法伤害的反应》。在《*2023*年*CHI*人机交互会议计算系统中的人为因素论文集》（*CHI '23*），*2023*年*4*月*23*日至*28*日，德国汉堡。ACM，美国纽约州纽约市，26页。https://doi.org/10.1145/3544548.3580953

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI '23, April 23–28, 2023, Hamburg, Germany
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3580953

1 引言

Artificial intelligence (AI) systems assist and make decisions in many high-risk scenarios (e.g., medical diagnostics [38], bail decision-making [4], hiring [109]), and their increasing use has raised new ethical concerns. Machines have been shown to discriminate against racial and gender minorities [28, 83] and even caused the deaths of their users [87]. Understanding how humans react to the harms caused by AI systems is critical for determining how these systems should be governed, developed, and adopted [14, 21]. Past research has shown that people underreact to algorithmic discrimination [13]. However, the opposite has also been reported, as evidenced by citizens attacking autonomous vehicles following the death of a pedestrian by a self-driving car [90]. These contradictory stories highlight the difficulty in comprehending people's collective reactions to algorithmic harm.

当前的研究正是由这一挑战所推动。我们借助责备的视角，考察人们[1] 对算法造成伤害的反应态度。有大量实证证据表明，当人工智能系统和机器人造成伤害时，人们会责备它们 [42, 56, 58, 69, 74], ，而许多哲学讨论则聚焦于机器是否适合作为责备的对象 [23, 100]。

We present three studies (N = 1,153 in total) that investigate how various factors influence laypeople's reactions to algorithmic harm. We first explore whether AI systems providing simple explanations about their decisions affects how much people blame them, their designers, and their users. Scholars have argued for explainable AI systems as a way to hold humans accountable when machines cause harm [53, 89]. In Study 1, we empirically test this proposal in the context of algorithmic decision-making and find that providing simple explanations has no effect on blame judgments. Nonetheless, blame directed at developers and users increase when the AI system presents a clearly discriminatory explanation for its decision (e.g., based on gender or race). The same effect does not occur in the case of machines, implying that people may underreact to algorithmic discrimination when it comes to assigning blame.

An exploratory analysis of Study 1 identified two factors that correlate with how people attribute blame to AI systems and other actors: 1) the perceived harmfulness of the decision and 2) the participants' attitudes towards AI systems. In Studies 2 and 3, we manipulate these factors and find that blame judgments of developers are largely shaped by the perceived fairness of the algorithmic decision, whereas blame directed at users is influenced more by its perceived harmfulness. A meta-analysis of the three studies

[1]我们用"反应性态度"一词来指人们对责备的判断。

indicate that fairness and harmfulness have little to no effect on blame towards machines, which is instead strongly associated with people's perceptions of AI systems as blameworthy agents.

开发者和用户都被高度归咎于算法危害，这表明存在一种共同责任框架，人们会将其反应性态度指向多个行为主体 [50]。我们讨论了普通人的预期可能如何影响未来的监管决策；监管决策应考虑到掌权者可能利用解释，将人们感知到的责任从自身转移出去 [71]。

Our findings demonstrate how laypeople's moral judgments of machines and humans differ. Even though fairness is a well-established factor in moral reasoning about human actors, it has no effect on blame judgments of AI systems. Instead, judgments about AI systems are determined by people's stance towards the possibility of blaming machines. Those who believe AI systems are not suitable recipients of blame choose not to blame them, whereas those who believe otherwise blame machines as much as other actors. We discuss how future decisions to integrate AI systems into the moral and social spheres will affect how people react to their actions.

2 背景

2.1 应受责备的人工智能系统?

Automated systems, such as AI and robots, are becoming more common in high-risk scenarios. For instance, they assist doctors prioritize patients who require urgent medical care and aid employers review job applicants to expedite the hiring process. They help judges make bail decisions and operate self-driving vehicles. These systems, however, are far from perfect. Decision-making algorithms have been found to discriminate against African American patients and defendants [4, 83]. There have been stories of AI recruiting tools disfavouring women in the hiring process [28] and self-driving cars killing pedestrians and passengers [87, 108]. This study looks into how laypeople, i.e., those subjected to AI systems, react to the harms these new technologies cause.

An extreme example of public reaction to harm caused by automated systems is the vandalism against autonomous vehicles following the killing of a pedestrian by a self-driving car [90]. According to Liu and Du [73], people overreact to crashes caused by autonomous vehicles, which could hinder their adoption. In contrast, other research suggests that people may underreact when algorithms make biased decisions compared to when a human is the decision-maker [12]. These contrasting examples highlight the complexity of understanding laypeople's reactions to algorithmic harm. Understanding how the public might react in such scenarios can help prevent future conflicts between the design of AI systems, their regulation, and public opinion [6, 14].

We study laypeople's reactions through the lens of blame. What it means to blame someone has been a contentious topic among moral philosophers. Scanlon [92] argues that judgments of blameworthiness are based on assessments of an actor's attitudes in relation to what is expected from them. In this view, to blame someone is to modify one's relationship with the blamee according to their perceived blameworthiness. Another perspective is put forward by Strawson [97], who suggests that blaming someone consists of a 一系列情绪反应，即所谓的"反应性态度"。Shoemaker [95] 主张，责备的作用是表明责备者对一套规范的认同，无论责备以何种方式付诸实施；而不是根据行为者的态度或行为来定义责备。

Alongside the theoretical discussions around blame, empirical research suggests that people blame machines for the harm they cause. When robots do not make a utilitarian decision, they are blamed more than their human counterparts [74].[2] Another study found that AI and human decision-makers are blamed similarly when it comes to bail decisions, despite the fact that automated systems were not given the same forward-looking responsibilities as their human counterparts [70].[3] Autonomous systems are blamed more than those that rely on human input, implying that people's perceptions of autonomy play a key role in how they react to algorithmic harm [42, 58]. Most research concludes that automated systems, such as AI systems, are more susceptible to blame than other machines (e.g., a vending machine) but not as much as humans [56].

Normative research has questioned the appropriateness and viability of individuals blaming automated systems. Scholars argue that blaming AI systems would be morally wrong because they are not appropriate subjects of retributive blame [27]. Machines lack the necessary moral understanding to be blamed [106] and do not possess the sentience required for comprehending what it means to harm someone [101]. These arguments concern whether AI systems and robots have the properties necessary to be blamed [23], and those who subscribe to this viewpoint agree that current automated systems do not.

In contrast, other scholars view blame as a social process that does not always track whether agents satisfy the necessary properties [100]. Instead, they suggest people react to wrongdoers before considering their sentience, autonomy, or any other property; these considerations can only serve as a secondary (and possibly mitigating) factor. Proponents of this viewpoint argue that individuals can adapt their reactive attitudes to machines, particularly if they appear to be blameworthy [25] or if doing so fulfills crucial social functions [96]. This perspective appears to be consistent with cognitive science studies [13, 74] and the literature on how people blame AI systems and robots (see above).

It is unclear whether people apply similar moral frameworks when judging machines and humans. For instance, past research suggests that whether a human agent is supervised impacts how much they are blamed [43, 49]; similarly, machines perceived as highly autonomous are blamed more than their supervised counterparts [42, 58, 69]. In contrast, it has also been shown that people may forgive humans but not machines for the same action [75, 93], suggesting differences in how people react to harm depending on the actor.

探讨机器造成伤害时人们会责怪谁至关重要，因为人们担心自主系统和

[2] Although we also mention that Lee et al. [66] provide a contrasting study, in which participants did not blame robots for their actions in the trolley dilemma.

[3] Backward-looking notions of responsibility (e.g., blame) pertain to previous actions, decisions, and consequences. Forward-looking responsibilities prescribe obligations concerning future actions [102]. Some suggest that not attending to an obligation (forward-looking responsibility) can lead to blame (backward-looking responsibility) [102].

自我学习系统会造成责任缺口 [77]。这一缺口认为，由于此类系统具备自主性和适应性，追究人类利益相关者的责任会变得困难。可责性是责任缺口的构成要素之一 [31], ，这引出了一个问题：当机器造成伤害时，是否有任何行为者适合作为责备对象。

Although distinct from the question of who should answer legally and morally for machine-cause harms, i.e., who is accountable, deciding who is to blame can help determine "contenders for the class of accountable actors," thus outlining who may face any potential sanctions for harmful outcomes [26]. This relationship between blame and accountability can be exemplified by the concern that AI systems may become responsibility shields by absorbing blame that should have been attributed to human agents [18, 54], exempting humans from the duty to answer for their actions and avoiding any penalties. This possibility highlights the importance of empirical research investigating how laypeople attribute blame to machines and human agents to ensure that accountability is not eroded when things go awry.

本研究探讨了三个可能影响人们在多大程度上责怪人工智能系统及其开发和部署者的因素：1）感知到的危害性，2）可解释性，以及3）个人对自动化系统的态度。

2.2 感知到的危害性

社会心理学理论认为，感知到的伤害是"道德的本质" [46]。然而，关于算法决策的感知有害性如何影响人们归咎于人工智能系统的程度，尚无研究加以探讨。Hidalgo 等人 [51] 发现，人类和机器实施的相同行为，可能被视为具有不同程度的伤害性。在本研究中，我们通过实验探究感知有害性如何影响责备判断。

2.3 可解释人工智能

我们在本研究中考察的另一个因素是可解释性。目前大多数人工智能系统都是黑箱，其输出通常难以被人类观察者理解。随着这些系统开始做出具有重大后果的决策，了解其运作方式变得至关重要。这一需求推动了可解释人工智能（XAI）领域的兴起，该领域致力于创建"能够提供细节或理由，使 [其]运作方式清晰或易于理解"的模型 [5]。

A review of the global guidelines for AI has found that explainability is the most prominent principle across all efforts to develop ethical AI systems [53]. Explainable machines have a wide range of functions, ranging from ensuring that systems are fair to evaluating whether they do what they are supposed to [65]. We focus on the proposal of explainability to ensure that human actors can maintain meaningful human control over automated systems and thus remain responsible for any harm they may cause [65, 89]. Explainable systems could fulfill legal requirements for holding developers and users responsible [11] and contribute to the perception that AI systems are *not* to blame and that human actors are the ones that should be held responsible. Explainability should thus help maintain developers and users as appropriate subjects of responsibility [65, 81, 89]. An early study seems to point to the相反，机器解释自己为何犯错，会减少用户受到的指责程度 [58]。

2.4 人们对人工智能的态度

The last factor we investigate concerns people's perceptions of and attitudes towards AI systems. Scholars argue that the public perception of new technologies has crucial implications for how they will be adopted, developed, and regulated [22]. Both accurate and misleading narratives about AI can drive and hinder public acceptance, as shown in the past with genetically modified (GM) crops, whose reception was shaped by public risk perceptions [21]. These attitudes can sway the design of machines and inform policymakers in both positive and negative directions [22]. Science fiction can also influence how people perceive these systems [22]; for instance, they may portray AI and robots as intelligent, sentient, and intentional, which should affect how people react to them. These portrayals could create false perceptions with moral and legal consequences to how laypeople interact with machines [17].

Recent research in HCI provides some evidence that people may have different stances towards and expectations about human and AI actors. For instance, people exhibit varying expectations about human- and machine-led decisions in different types of decision-making tasks [63]. AI systems' decisions are perceived as less fair and trustworthy than identical human decisions in settings that require human skills (i.e., subjective tasks, such as hiring), but not in settings perceived as mechanical (i.e., objective tasks, such as work scheduling) [20, 67]. People exhibit a preference for humans over AI systems in high-stakes scenarios [64] and in settings associated with higher degrees of inherent uncertainty [32], such as investing and medical decision-making. Past work indicates that making an algorithm transparent to users influences their views towards AI recommendations and decisions [15, 37], suggesting that the provenance of explanations can also impact lay perceptions of algorithmic decision-making (as explored in this research). Considering people's different expectations about humans and AI systems, we thus inquire whether lay perceptions of AI systems and their decisions influence how much they (and other actors) are blamed.

Lay perceptions of AI systems also differ based on one's understanding of AI. People have diverse mental models of AI systems, which are different across groups and over time, impacting how they perceive and use algorithms [35, 36]. In this work, we do not attempt to create a standardized understanding of decision-making AI across participants (e.g., by defining it). Doing so would not reflect the real world [57] but instead limit our efforts to capture and understand how people attribute blame when AI systems cause harm in the real world. In contrast, as shown below, we explore how these different lay attitudes towards AI systems (and the possibility of blaming them) help determine whether machines deserve blame.

2.5 研究预备知识

In the present research, we focus on AI systems that make consequential decisions. Algorithms have been developed to assist humans making decisions in a wide range of scenarios (e.g., [4, 28]), and regulatory proposals defend that AI systems should only be

deployed with human oversight [105]. Completely delegating decisions to algorithms may be illegal in certain domains, such as bail decisions, in which defendants have the right to be heard by a competent human judge [47]. We note, however, that several AI systems currently take on the role of decision-makers in the real world. Existing algorithms decide who is automatically enrolled in health care programs [83] and determine which job postings are shown to online users, impacting their job prospects [29]. Decision-making AI systems can also decide which images are returned by search results, which research has found to reinforce racial and gender stereotypes and shape people's views of their job opportunities [80].

Even in scenarios where AI systems are deployed alongside humans, they may still be decision-makers in practice. Resume-screening algorithms decide which job applications should be evaluated by a human recruiter, rejecting an initial pool of applicants without human oversight [109]. In scenarios where AI systems rank applicants instead of directly rejecting them, algorithms can still practically decide who will be evaluated by humans by determining who should appear on the first few result pages. Tenant-screening algorithms are marketed as decision-making products and disqualify applicants "without providing any details [...] that would have allowed the property manager to make his own decision" [59], obscuring human decision-making. Humans tend to overrely on machines [62] and decision-making algorithms [8], particularly if they are explainable [9] as the ones explored in this research. Hence, AI systems can shape people's decision-making processes and limit their courses of action, taking on the role of a real decision-maker.

Finally, we clarify how we operationalize "blame," which lacks a unified definition. We neither emphasize nor ask that our study participants adhere to a particular definition. Instead, we collect lay intuitions about blame, allowing participants to interpret what blame means and disclose their judgments accordingly. These judgments are referred to as *"blame judgments"* in this work. Past research followed a similar approach to explore how people blame machines (e.g., [75, 76]) and other human actors (see a review of the moral psychology literature on blame [74]). Had we provided a particular definition of blame to participants, we would have biased their responses and measured how well they could follow the desired interpretation of blame. In contrast, our decision allowed us to explore how laypeople attribute blame to AI systems compared to human agents, expanding on the extensive literature investigating blame judgments of humans. As we will show below, lay judgments of AI systems have unique features that are not considered when blaming human agents, providing insights that could inform future research aiming to define what blame is and who (or what) can be blamed (see Section 7).

3 对人工智能系统、开发者和用户的责备判断

我们呈现三项研究，探讨人们对算法决策可解释性和危害性的感知，以及人们对人工智能系统的态度，如何影响普通人对机器造成伤害的反应。基于第 2 节介绍的既有研究，我们提出三个研究问题：

RQ1）事后解释如何影响人们对 AI 系统、其开发者和用户的责任归咎判断?

RQ2）对算法决策有害程度的感知如何影响人们对 AI 系统、其开发者和用户的责备判断？RQ3）人们对 AI 系统自主性的看法如何影响其对 AI 系统、其开发者和用户的责备判断?

Study 1 tested the effect of post-hoc explanations on blame judgments (RQ1) and identified two additional factors that correlated with how blame was distributed: perceived harmfulness and participants' attitudes towards AI systems. Study 2 investigated how perceived harmfulness influences people's judgments (RQ2), whereas Study 3 explored whether the perceived autonomy of AI systems is the main driver of the blame directed at them (RQ3). Study 3 was not successful in manipulating perceived autonomy; hence, we conducted a qualitative assessment of participants' open-ended justifications and found people's attitudes towards the possibility of blaming machines to be the most significant factor determining whether people blame AI systems. Figure 1 presents a high-level overview of the three studies, their treatment conditions, and measures. All studies were approved by the first author's IRB, and we make our data and scripts available at https://bit.ly/3DtFH7H.

3.1 研究1

研究1的主要目的是通过实验检验可解释性如何影响人们对人工智能系统及其开发者和用户的归责（RQ1）。本研究设计受到学者观点的启发：能够为其决策提供解释的人工智能系统，会使人们认为应当归咎于人类行为者。我们还进行了探索性分析，考察感知到的危害性与普通人判断之间的相关性，并对人们解释其判断的开放式回答进行了定性分析。

3.1.1 Study Design. After agreeing to the research terms, participants were shown a three sentences-long introduction that explained how AI systems are currently used to make and assist decisions in a wide range of environments. Participants were then shown a vignette in which Systemy, a local technology firm, used an AI system to hire new software developers. The scenario presented Taylor, a junior software developer that applied for the position and was later rejected by the AI system. The AI system explained its decision on a between-subjects basis: it either did not give any explanation (*no explanation/none*), explained that Taylor did not have the necessary experience (*explanation*), or justified its decision based on gender (i.e., because Taylor is a woman; *discriminatory explanation*).

A few design choices are worth clarifying. The hiring domain was selected due to the increasing use of decision-making systems in selection processes (e.g., [109]). We focused on gender-based discrimination, being inspired by recent incidents of hiring AI systems discriminating against women [28]. The phrasing of the vignette was inspired by the work of Plane et al. [88]. We named the victim Taylor—a unisex name that is common among both White and African-Americans—to minimize the salience of the victim's gender and race in the treatments in which these features are not a part of the experimental manipulation.

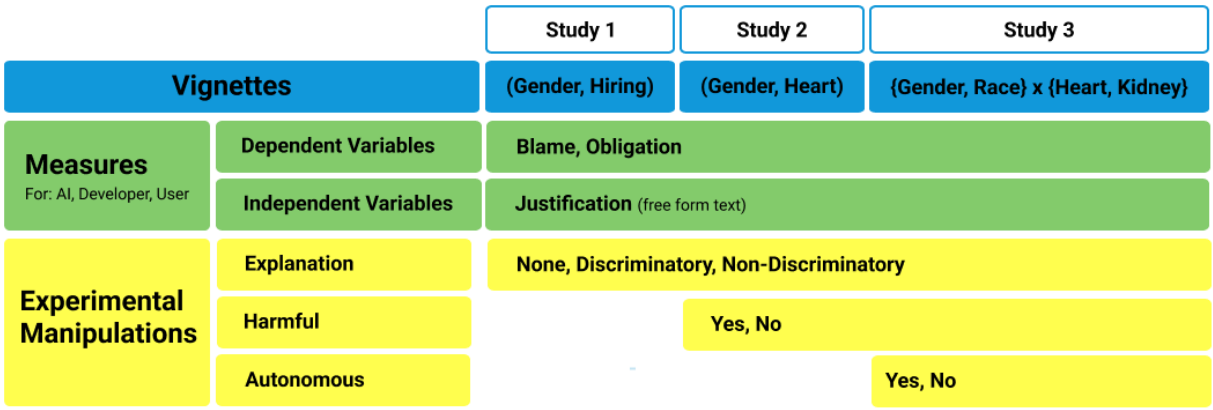

图1：研究1、2和3所用方法的总体概览。具体情境描述和问题措辞见附录。

Our treatment conditions were inspired by the natural language explanations employed by Facebook when justifying which advertisements are shown to its users [3]. We employed simple text-based explanations to mitigate the effect of any confounding variables by presenting justifications that could be understood by any participant. The development of explainable algorithms aims to make them understandable to an audience, be it the general public, policymakers, or developers [5]. Hence, our study provided explanations that would be understandable to laypeople, i.e., the study's audience, by removing any complexity that could impact their interpretation. This design choice allowed us to measure the effect of explainability on blame judgments while mitigating the influence of latent variables, such as their ability to comprehend specific types of explanations.

Our study also investigated how discriminatory explanations impact blame judgments. We employed illegal explanations to ensure that our treatment condition was viewed as discriminatory.[4] By presenting egregious justifications based on gender, we mitigated the effect of confounding variables that could have influenced people's judgments of fairness and blame. We acknowledge that most existing algorithms that were found to discriminate based on gender did so indirectly since they did not directly consider gender in their decision-making processes (e.g., [28, 29]). Hence, to ensure that our results could be extended to more realistic scenarios, we replicated our results with legal but clearly discriminatory explanations in Study 3.

3.1.2 Measures. Participants were first asked three questions: 1) how much *blame* the AI system, its developer, and Systemy (i.e., its user) deserved for the decision not to hire Taylor and 2) how responsible these actors were for ensuring that the decision was correct, i.e., to what extent they prescribed an *obligation* to each of the actors. Participants also 3) explained their judgments in a free text form. These three questions were shown separately for each

[4]在美国及许多其他司法管辖区，承认招聘决定是基于性别的属于违法行为。这并不是说就业市场中不存在基于性别的歧视；然而，为了避免遭到法律追诉，决策者不会给出如此明确的歧视性解释。

行动者的呈现顺序在参与者之间是随机的。

随后，受访者表示了 AI系统在多大程度上解释了其决策（感知可解释性），以及该决策在多大程度上是公平的（感知公平性）和有害的（感知有害性）。所有问题均采用从0 到6 编码的 7 点量表进行回答。我们在附录中呈现了所有测量项目和材料。调查最后包括一系列人口统计学问题和事后说明。

3.1.3 Participants. We conducted a power analysis to calculate the minimum sample size. A Wilcoxon-Mann-Whitney two-tailed test requires 67 respondents per treatment group to detect an effect size of Cohen's $d = 0.5$ at the significance level of 0.05 with 0.8 power. Hence, we recruited 212 participants through the Prolific crowdsourcing platform [85] to account for attention-check failures, i.e., at least 67 participants per treatment group, assuming a 5% attention-check failure rate. We targeted US residents that had previously completed at least 50 tasks on Prolific, with an approval rate of 95% or above. Three participants failed at least one of the attention check questions that were presented before and after the vignette, resulting in a final sample size of 209 (47.85% women; $M_{age} = 36.08$, $SD_{age} = 12.19$). All participants were compensated a median of US$7.30 per hour.

3.1.4 Results. We employed one-way analysis of variance (ANOVA) tests to identify differences between treatment conditions. Decisions followed by a discriminatory explanation were perceived as more unfair than those complemented by a justification based on the applicant's experience or no explanation at all ($F(2,206) = 50.15$, $p < .001$, $\eta_p^2 = 0.33$). Explainable systems were also regarded as more explainable than their opaque counterparts regardless of whether their explanations were discriminatory or not ($F(2,206) = 24.89$, $p < .001$, $\eta_p^2 = 0.19$). These results show that the between-subjects conditions achieved the desired effect on perceived explainability and fairness (see Appendix for mean values and post-hoc pairwise tests).

Figure 2 presents participants' blame judgments of the AI system, its developer, and user (see Appendix for numerical values and

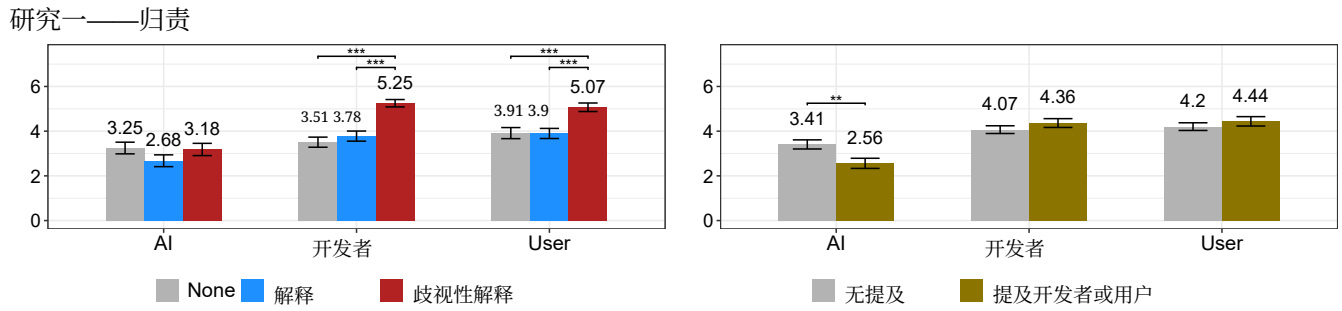

Figure 2: Blame judgments of the AI system, its developer, and user involved in making hiring decisions in Study 1. Participants were randomly assigned to a treatment condition where the AI system either did not provide any explanation for its decision (i.e., *None*), justified its reasoning based on the applicant's experience (i.e., *Explanation*), or discriminated against the applicant based on her gender (i.e., *Discr. Explanation*). Participants were also categorized into those that explained their blame judgments of the AI system with *mentions to its developer or user* and those who did not highlight the role of the human actors developing and deploying the AI system (i.e., *no mentions*). Standard errors are presented as error bars. $^{*}p$ <.05, $^{**}p$ <.01, $^{***}p$ <.001.

Tukey's HSD post-hoc tests). Explainability did not influence the extent to which the AI system was blamed ($F(2,206) = 1.35$, $p = .262$, $\eta_p^2 = 0.01$). In contrast, the developer of a discriminatory AI system was blamed more than one that developed a non-discriminatory explainable system or an opaque decision-maker ($F(2,206) = 21.01$, $p < .001$, $\eta_p^2 = 0.17$). Similarly, participants blamed the user of a discriminatory AI system more than other users ($F(2,206) = 9.30$, $p < .001$, $\eta_p^2 = 0.08$).

参与者对所有行为主体的义务归因并未受到研究处理的影响（$p > .05$ 对所有行为主体均如此）。尽管如此，与其开发者（$M = 4.10$, $SD = 1.96$）或用户（$M = 4.03$, $SD = 2.07$）相比，AI系统（$M = 2.66$, $SD = 2.16$）被赋予义务的程度较低。

An exploratory analysis suggested that perceived harmfulness correlated with blame judgments of all actors, such that participants who perceived the decision as more harmful blamed the AI system ($r = 0.20$, $p < .005$), its developer ($r = 0.49$, $p < .001$), and user ($r = 0.35$, $p < .001$) more. In contrast, perceived harmfulness was not associated with attribution of obligations to any of the actors (all $p > .05$).

Investigating participants' open-ended justifications for their blame judgments of the AI system suggested that some rationalized their reactive attitudes based on the perception that these systems are programmed and used by humans, and thus should not be blamed. To identify those who mentioned the AI system's developer or user in their justifications, we used the regular expression *"systemy|dev|progr|company|user|medical|hospital|west|doct"* to identify participants that highlighted the role of developers or users in their blame judgments of the AI system. This regular expression was crafted after a qualitative analysis of participants' open-ended responses and aimed to identify those who justified their blame judgments of machines with mentions to human and collective agents (e.g., the programmer, Systemy). This regular expression was iteratively updated throughout the three studies presented in this paper and contains terms used to identify the AI system's user in Studies 2 and 3. We categorized participants into two groups: 在解释中提及 AI 系统开发者或使用者的人（43.54%）与未提及的人（56.46%）。

We used a 3 (treatment) x 2 (mentions or not) ANOVA to account for this variable (see Figure 2). We found that participants who mentioned the developer or user blamed the AI system less than those who did not ($F(1,203) = 7.52$, $p < .01$, $\eta_p^2 = 0.04$). In contrast, this variable did not influence blame judgments of developers ($F(1,203) = 0.32$, $p = .574$, $\eta_p^2 = 0.00$) and users ($F(1,203) = 0.26$, $p = .610$, $\eta_p^2 = 0.00$). The main effects of the explainability treatment were consistent with our initial analysis, and we did not observe any significant interaction between these two factors (all $p > .05$)

3.1.5 Discussion. **Explainability, by itself, did not impact the extent to which the AI system, its developer, and user were blamed. However, the perceived fairness of the AI system's explanation did.** The developers and users of an AI system that provided a discriminatory explanation received more blame, while reactive attitudes directed at the machine did not change accordingly. Surprisingly, blame towards the AI system did not increase when it provided an egregious and unfair explanation based on gender. With respect to explainability, explainable AI systems were not blamed to a larger extent than their opaque counterparts. Their developers and users were also blamed similarly to those of a non-explainable system. This result contrasts with the proposal of explainability to facilitate identifying which human agent is responsible when machines cause harm.

An exploratory analysis of Study 1 suggested that perceived harmfulness might influence blame judgments of machines and other actors differently. However, our scenario was not perceived as particularly harmful; mean perceived harmfulness was close to the mid-point ($M = 3.86$, $SD = 1.81$). A possible explanation is that the rejection of a job application might have been viewed as an opportunity loss rather than a clearly harmful outcome. Moreover, the design of Study 1 did not control for the interaction between perceived fairness and harmfulness, preventing us from distinguishing their distinct effects on blame judgments. We examine this factor further in Studies 2 and 3.

Nevertheless, our results suggest that the main factor at play in blame judgments of machines may not be perceived harmfulness or fairness but people's perceptions of these systems. Those who justified their blame judgments by mentioning the developer or user tended to blame the machine less than those who did not—a finding we explore further in subsequent studies and Section 5.

最后，我们发现，相较于开发者和用户，人们较少认为人工智能系统负有义务（例如面向未来的责任），不过这一点在不同处理条件之间并无差异。这些结果与先前的研究一致：研究表明，与人类相比，人们较少认为机器负有面向未来的责任[70]，而且这类判断不受行为后果的影响[102]。

3.2 研究二

研究2通过实验考察了感知到的有害性如何影响人们对机器及参与算法决策的人类行为者的责备判断（RQ2）。研究1中对有害性的判断与感知到的公平性相关；研究2将这两个因素区分开来，并探究它们如何在影响人们对人工智能系统造成的伤害的反应时相加或相互作用。

We also complemented our research with a different domain: medical decision-making. Medicine has received considerable attention from AI research, with algorithms being used to discover new drugs [40], identify patients at risk [83], and many other functions. We employed a vignette depicting disparities in the medical treatment of men and women [84], specifically in terms of the diagnosis and treatment of heart attacks. Coronary heart disease is underdiagnosed and undertreated in women [79], and women are less likely to survive an acute myocardial infarction (commonly known as a heart attack), partially due to the differences in the clinical presentation [110] and the medical treatment they receive [24]. In contrast to the vignette employed in Study 1, the medical scenario allowed us to vary perceived harmfulness experimentally.

3.2.1 Study Design. Participants took part in a study similar to Study 1. Study 2 introduced West Medical, a local hospital that uses an AI system to assess the condition of emergency room patients. The scenario then presented Taylor, a woman that went to West Medical's emergency room after feeling discomfort in her left arm and nausea for three days straight, i.e., common symptoms of a heart attack. Taylor was classified as a non-critical patient by the AI system and either received no justification (*no explanation/none*), was told that she was classified as such because she did not have any pre-existing conditions (*explanation*), or was informed that this classification was due to her gender (i.e., because Taylor is a woman; *discriminatory explanation*). Similarly to Study 1, we employed egregious explanations based on gender to mitigate the effect of confounding variables in the discriminatory explanation condition.

研究2还改变了AI系统决策的结果。等待八小时后，Taylor要么接受了医生的检查并被开具了一些药物（无害），要么突发心脏病，在未获得适当医疗救治的情况下死亡（有害）。所有处理条件均在被试间随机分配，因此每位受访者阅读了 3（可解释性处理）× 2（有害性处理）情景中的一个。

3.2.2 测量。参与者回答了与研究1相同的一组问题。所有问题均采用7点量表作答，编码范围为0至6。详情见附录。

3.2.3 Participants. Considering the same power analysis from Study 1, which required 67 respondents per treatment group, we recruited 410 participants through Prolific. None of the participants had participated in Study 1. We employed the same recruitment and exclusion criteria as in Study 1. Four participants failed the attention check, resulting in a final sample size of 406 (46.55% women; M_{age} = 34.37, SD_{age} = 11.81). All participants were compensated a median of US$7.66 per hour.

3.2.4 结果。我们采用了 3（可解释性处理）× 2（危害性处理）的方差分析模型。附录中包含对感知可解释性、公平性和危害性的分析，作为操纵检验。我们的实验操纵是有效的：人们认为可解释机器更具可解释性，认为歧视性决策更不公平，并认为有害后果危害更大。

图3 展示了研究2中按处理条件划分的对AI 系统、其开发者和用户的责备判断。两种处理条件之间的交互项在对所有行为主体的责备判断中均不显著（p >.05）。因此，我们呈现不受任何交互效应影响的均值和标准误。附录包含描述性统计和 Tukey 的 HSD 检验结果。

Explainability had a small effect on judgments of the AI system: a machine that explained its decision without explicit discrimination was blamed marginally less than its discriminatory counterpart (F(2,400) = 3.88, p <.05, η_p^2 = 0.02). An AI system whose decision led to the death of the patient was blamed more (F(1,400) = 8.54, p <.005, η_p^2 = 0.02).

Blame directed at the AI system's developer was highest for that of a discriminatory machine (F(2,400) = 13.31, p <.001, η_p^2 = 0.06). The effect of harmfulness was also significant: the developer of a system that led to the death of the patient received more blame (F(1,400) = 9.25, p <.005, η_p^2 = 0.02).

Harmfulness was a more significant factor in blame judgments of the AI system's user (i.e., West Medical; F(1,400) = 27.98, p <.001, η_p^2 = 0.07). The user of a machine whose decision led to the patient's death was blamed more. Explainability influenced reactive attitudes towards the user to a lesser degree (F(2,400) = 4.24, p <.05, η_p^2 = 0.02), such that users of a discriminatory system received more blame than those of a non-discriminatory explainable algorithm.

None of the treatment conditions influenced participants' attribution of obligations to the AI system, its developer, and user (p >.05 for all main effects). This finding is consistent with Study 1. The only significant yet small effect was observed in the interaction between explainability and harmfulness in attributions of obligations to the user (F(2,400) = 3.97, p <.05, η_p^2 = 0.02). Overall, the AI system (M = 3.26, SD = 2.26) was ascribed forward-looking responsibilities to a lesser extent than its developer (M = 4.40, SD = 1.78) and user (M = 4.71, SD = 1.70).

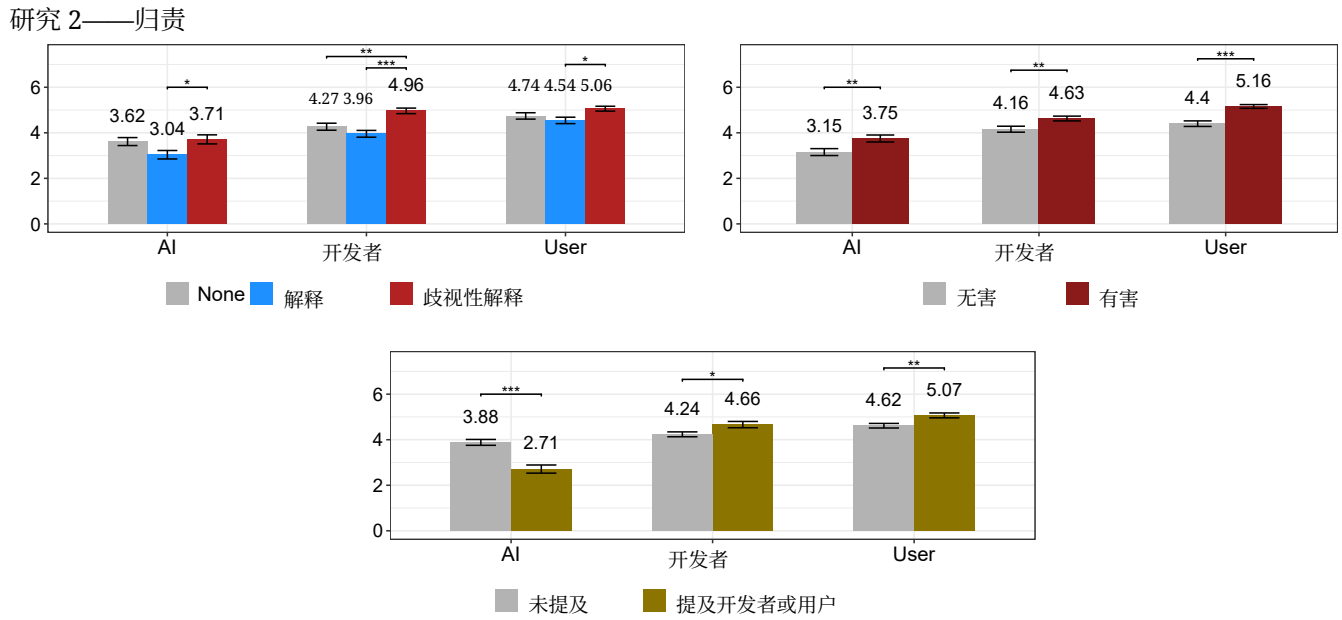

Figure 3: Blame judgments of the AI system, its developer, and user involved in making medical prioritization decisions in Study 2. Participants were randomly assigned to a treatment condition where the AI system either did not provide any explanation for its decision (i.e., *None*), justified its reasoning based on the patient's medical pre-conditions (i.e., *Explanation*), or discriminated against the patient based on her gender (i.e., *Discr. Explanation*). The vignette concluded with either the patient receiving medication for her symptoms (i.e., *Not Harmful*) or dying without receiving proper medical care (i.e., *Harmful*). Participants were also categorized into those that explained their blame judgments of the AI system with *mentions to its developer or user* and those who did not highlight the role of the human actors developing and deploying the AI system (i.e., *no mentions*). Standard errors are presented as error bars. $^{*}p$ <.05, $^{**}p$ <.01, $^{***}p$ <.001.

Similarly to Study 1, we examined participants' justifications for their blame judgments of the AI system. Using the same categorization method as in Study 1, 36.70% of participants mentioned the developer or user when explaining their judgments. We employed 3 (explainability) x 2 (harmfulness) x 2 (mentions to developer or user or not) ANOVA models to identify the effect of this variable on how people blame all actors. The AI system was blamed significantly more when participants did *not* mention its developer or user ($F(1,394) = 32.77$, p <.001, $\eta_p^2 = 0.08$). In contrast, the AI system's developer received more blame when participants mentioned the developer or user ($F(1,394) = 4.44$, p <.05, $\eta_p^2 = 0.01$). The user was also blamed more when mentioned ($F(1,394) = 6.85$, p <.01, $\eta_p^2 = 0.02$). The effect of explainability and harmfulness on judgments of all actors did not change from our initial analyses.

3.2.5 Discussion. **Both explainability and harmfulness influenced how much participants blamed the AI system's developer and user. As in Study 1, however, it was not explainability itself that increased blame directed at these entities; instead, it was the perceived fairness of the AI system's explanation. Importantly, the effect size of each treatment condition differed by actor.** Blame directed at the AI system's developer was influenced the most by fairness considerations ($\eta_p^2 = 0.06$, i.e., a medium effect size). In contrast, reactive attitudes towards the user were mostly determined by the perceived harmfulness of 这一决定（$\eta_p^2 = 0.07$）。这些结果表明，责备可能部分取决于行为者对决策算法所拥有的控制程度。开发者可以控制人工智能系统中实施哪些解释，而用户只能决定是否采用该系统，以及它可能如何影响那些受到其影响的人。

Both explainability and harmfulness showed only a small effect on how participants blamed the AI system ($\eta_p^2 = 0.02$, i.e., a small effect size). Once more, the AI system was not blamed to a larger degree after justifying its decision with a clearly illegal and unfair explanation. As in Study 1, participants seem to judge machines differently from other actors. **Reactive attitudes towards the AI system were associated with how participants justified their judgments ($\eta_p^2 = 0.07$); the AI system was blamed depending on the perceived roles of its developer and user.** Investigating participants' open-ended responses suggested that participants who view the AI system as constrained by its programming blame it less, a hypothesis we explored in Study 3.

3.3 研究三

Study 3 had two major objectives. The first was to examine the effect of the AI system's perceived autonomy on how people react to algorithmic harm (RQ3). We test the hypothesis that people's judgments of AI systems are influenced by the perception that these machines' behavior is restricted by their programming, as

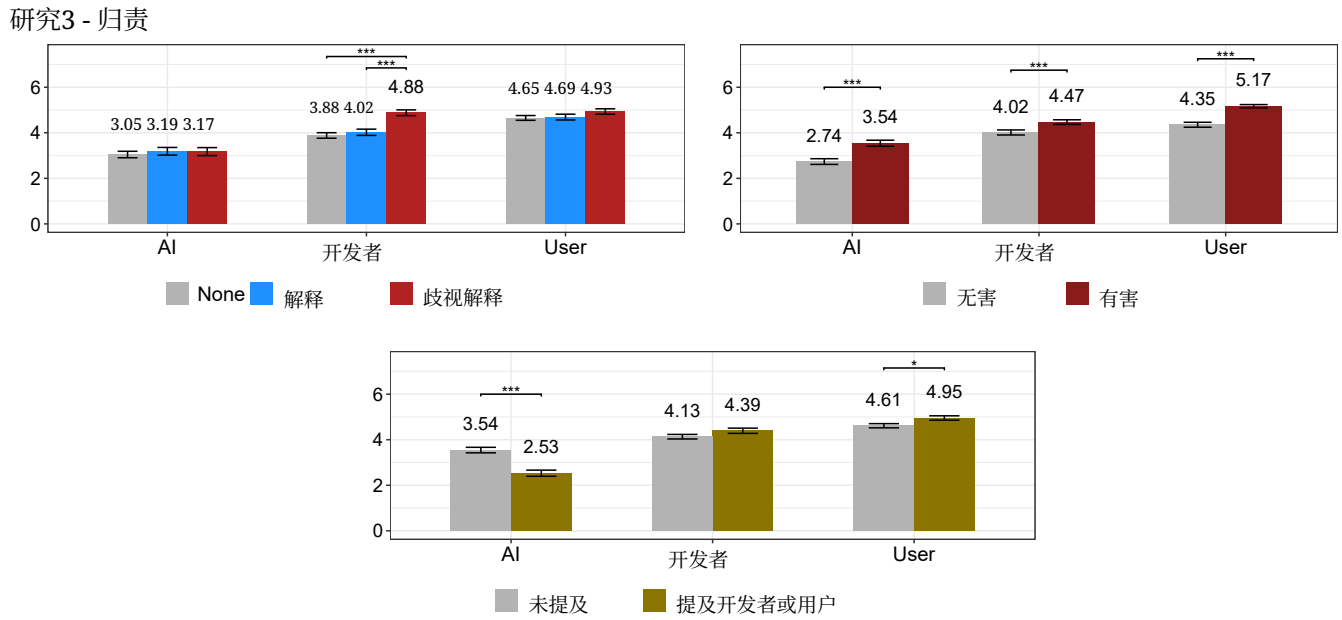

Figure 4: Blame judgments of the AI system, its developer, and user involved in making medical prioritization decisions in Study 3. Participants were randomly assigned to a treatment condition where the AI system either did not provide any explanation for its decision (i.e., *None*), justified its reasoning based on the patient's medical pre-conditions (i.e., *Explanation*), or discriminated against the patient based on their race or gender (i.e., *Discr. Explanation*). The vignette concluded with either the patient receiving medication for their symptoms (i.e., *Not Harmful*) or dying without receiving proper medical care (i.e., *Harmful*). Participants were also categorized into those that explained their blame judgments of the AI system with *mentions to its developer or user* and those who did not highlight the role of the human actors developing and deploying the AI system (i.e., *no mentions*). Standard errors are presented as error bars. $^{*}p$ <.05, $^{**}p$ <.01, $^{***}p$ <.001.

suggested by Study 2 and prior work [42, 58]. The second objective was to replicate our findings using different vignettes in the medical domain. The discriminatory explanations presented in Studies 1 and 2 were not realistic as they justify the decisions by describing procedures that are largely illegal in the real world—namely, direct discrimination based on legally protected features. Hence, Study 3 additionally employed discriminatory explanations that could be reasonably adopted in the medical domain.

3.3.1 Study Design. Participants took part in a study similar to Study 2 but with two major modifications. First, participants were randomly assigned to a vignette that varied with respect to both the characteristics of the patient looking for medical assistance and the condition from which the patient was suffering. We employed 2 x 2 vignettes that introduced Taylor, who was either a woman or an African American person, suffering from symptoms suggesting either a heart attack or kidney failure. In addition to gender-based discrimination, Study 3 considered racial discrimination in the medical domain [1]. We added vignettes related to kidney failure, inspired by research that has found that African American people are less likely to receive optimal treatment for chronic kidney disease and are more likely to progress to kidney failure [82]. In contrast to the explanations based on gender used in Studies 1 and 2, which would not be used in practice, race may indeed be used as a variable in medical decisions related to kidney 疾病 [2]。这些情境短文还纳入了我们此前对可解释性和危害性的处理，并作了相应调整。例如，如果将泰勒介绍为非裔美国人，那么泰勒就会因种族而受到歧视。参与者被随机分配到四个情境短文之一。

Second, we included an additional treatment condition. After introducing West Medical and its usage of an AI system for patient prioritization, some participants were told that the AI system learned which patients to prioritize without explicit rules set by its developer (*autonomous*). Other participants were not shown any information about how the AI system learned how to make decisions (*N/A*); this treatment sought to understand how people's conception that these systems are under the control of their developers influences how machines are blamed. Hence, participants were randomly assigned between-subjects to one of 3 (explainability) x 2 (harmfulness) x 2 (autonomy) treatment conditions.

3.3.2 Measures. Participants answered the same set of questions from Study 2. In addition, participants indicated the extent to which they thought "the AI system's decision was under the control of its developer," "the AI system was programmed to make this decision," and "the AI system's decision was explicitly programmed by its developer." These questions were used as manipulation checks for our autonomy treatment condition. All questions were answered

using a 7-point scale coded from 0 to 6. The Appendix presents all measures and materials.

3.3.3 Participants. We recruited 550 participants via Prolific using the same recruitment and exclusion criteria as earlier, out of whom 12 failed the attention check. No participant had participated in the previous studies. Our final sample comprised 538 participants (45.91% women; M_{age} = 35.53, SD_{age} = 11.71), achieving a power of 0.85 to detect the smallest effect size we had found in previous studies (η_p^2 = 0.02) at a significance level of 0.05. All participants were compensated at a median rate of US$10.35 per hour.

3.3.4 Results. We employed 3 (explainability) x 2 (harmfulness) x 2 (autonomy) ANOVA models and present a comprehensive manipulation check analysis in the Appendix. Our explainability, fairness, and harmfulness manipulations obtained the desired effect. Perceived fairness did not differ between the illegal scenarios, in which Taylor was discriminated against based on gender, and the lawful vignettes, in which race was given as an explanation (F(3,523) = 1.56, p = .199, η_p^2 = 0.01). However, the newly proposed autonomy treatment did not obtain the desired effect; the condition did not significantly increase perceived autonomy (average of responses to the three autonomy-related questions, Cronbach's α = 0.82; F(1,523) = 6.00, p <.05, η_p^2 = 0.01). An initial analysis of blame judgments did not suggest a significant main effect of the autonomy manipulation factor (all p >.05).

Hence, we discarded the autonomy treatment for all subsequent analyses as it did not achieve the desired effect. Adding it to our analysis did not modify our results. We present findings from 3 (explainability) x 2 (harmfulness) x 2 (mentions to developer or user—40.71% of participants—or not) ANOVA models, as done in Study 2. Analyses of participants' judgments using a 2 (explainability) x 2 (harmfulness) ANOVA model were indistinguishable from the reported results. We also include the vignette shown to participants as a fixed effect in all models.

图4呈现了研究3中参与者的归责判断（描述性统计和Tukey HSD事后检验见附录）。总体而言，我们的结果与研究1和研究2的发现一致。处理条件之间的一些交互项仅达到边缘显著水平（$\eta_p^2 \leq .01$）。我们独立于这些交互效应呈现均值。

Participants that justified their judgments of the AI system with mentions to its developer or user blamed the machine less than those who did not (F(1,523) = 31.28, p <.001, η_p^2 = 0.06). An AI system whose decision led to the death of a patient was blamed more than its counterpart (F(1,523) = 21.00, p <.001, η_p^2 = 0.04). Explainability did not influence people's reactive attitudes towards the AI system (F(2,523) = 0.82, p = .443, η_p^2 = 0.00).

The developer of a discriminatory AI system was blamed more than that of a non-discriminatory explainable machine or opaque system (F(2,523) = 20.45, p <.001, η_p^2 = 0.07). Harmfulness also influenced the extent to which the AI system's developer was blamed (F(1,523) = 13.64, p <.001, η_p^2 = 0.03). The developer of a machine that led to the death of the patient received more blame. Whether participants mentioned the AI system's developer or user when justifying their judgments of machines did not impact participants' 对开发者的反应性态度（F(1,523) = 1.80, p = .180, η_p^2 = 0.00）。

The main factor at play in judgments of the AI system's user was the decision's perceived harmfulness (F(1,523) = 43.46, p <.001, η_p^2 = 0.08). Users of a system that led to the death of the patient were blamed more than their counterparts. We also observed a small effect of people's justification of their blame judgments of machines (F(1,523) = 5.58, p <.05, η_p^2 = 0.01), such that those who mentioned the developer or user blamed the user more. Explainability had a marginal effect on people's judgments (F(2,523) = 3.28, p <.05, η_p^2 = 0.01), as did the vignette (F(3,523) = 2.99, p <.05, η_p^2 = 0.02).

We observed a series of small effects in people's attribution of obligations to the AI system, its developer, and user (all η_p^2 < 0.03). Overall, the AI system was deemed less responsible in a forward-looking manner (M = 3.27, SD = 2.10) than its developer (M = 4.33, SD = 1.72) and user (M = 4.93, SD = 1.51).

3.3.5 Discussion. Study 3 replicated our previous findings concerning the effect of explainability, fairness, and harmfulness on blame judgments of AI systems, their designers, and users in a wider range of vignettes. As in previous studies, explainability alone did not increase blame directed at any of the entities, which is in contrast to the proposal of explainability to facilitate identifying responsible human actors. **Nonetheless, our results demonstrate how different moral considerations shape blame directed at developers and users distinctively. Fairness judgments are more strongly associated with blame towards the AI system's developer, whereas judgments of users are more influenced by perceived harmfulness.**

感知到的公平性并未影响人们对人工智能系统的反应性态度。这些结果适用于研究1至3中采用的恶劣解释，以及研究3中基于种族的合法但具有歧视性的辩解。尽管感知到的有害性似乎会影响人们对人工智能系统的责备，但其影响较小，并且受人们对这些系统的认知所主导。

From Study 2's findings, we hypothesized that people's belief that the programming of AI systems constrains their behavior would be the main determinant of their blame judgments. We thus attempted to experimentally manipulate the perceived autonomy of the AI system. However, the treatment did not significantly impact people's perceptions, suggesting that these beliefs may not be easily mutable. This negative result does not imply that perceived autonomy does not impact blame judgments, which would challenge prior work [42, 58]. To better understand what specific attitudes may influence blame towards AI systems, we expand our analysis of how people justify their judgments of machines in Section 5.

4 元分析

Following recent best-practices recommendations [78], we present a within-paper meta-analysis to explore the average effect size of our treatment conditions across the three studies. Researchers caution against relying on single studies when evaluating the robustness and reliability of an effect, and an internal meta-analysis can help synthesize findings from multiple studies [44]. Hence, we conducted a meta-analysis of blame judgments using the *metacont* function from the *meta* package for the R programming language [94]. The meta-analysis calculates the treatments' weighted average effect

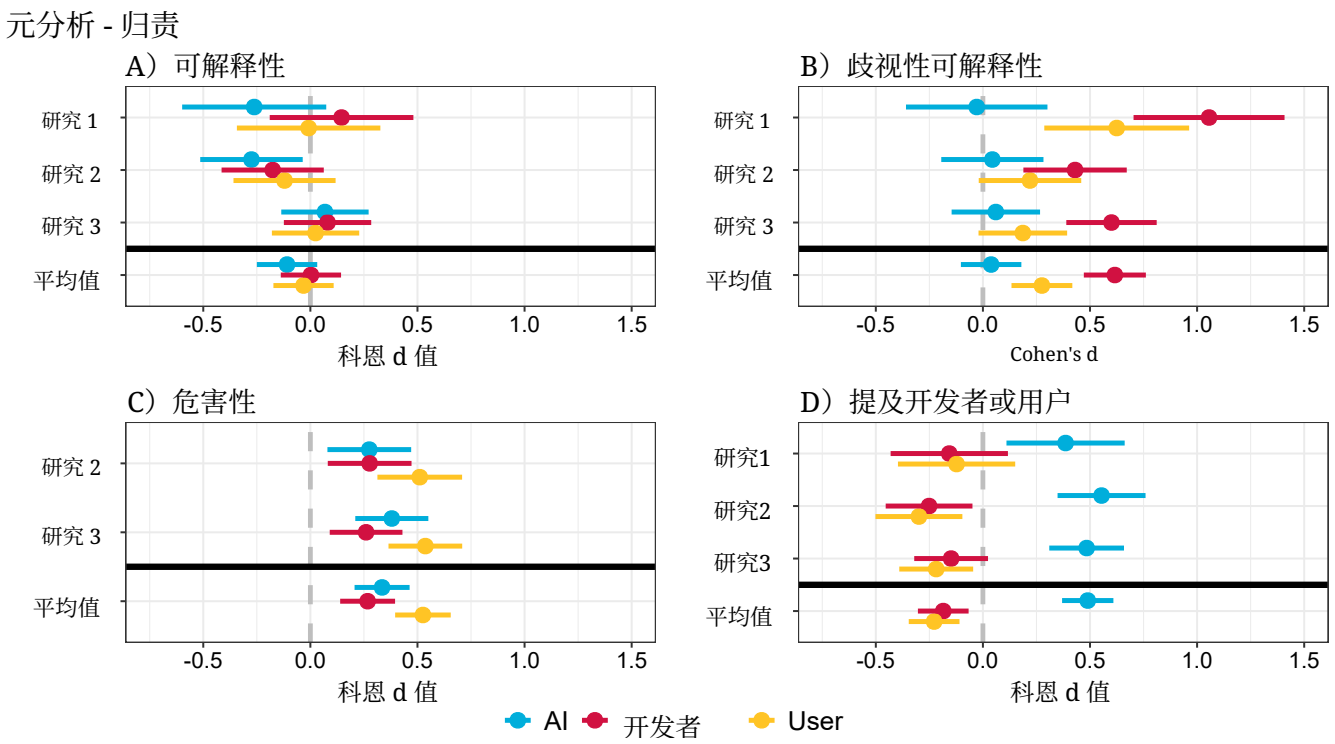

图5：对人工智能系统、其设计者及用户的责备判断的元分析。我们呈现了所有研究与变量中处理条件与对照条件之间的效应量（Cohen's *d*）。我们使用了固定效应元分析模型，并以误差线呈现95%置信区间。

所有研究中的效应量（Cohen's *d*），并在图 5 中与各项研究的估计值一并呈现。

Whether the AI system was explainable did not impact judgments of any of the actors (Figure 5A; RQ1); in contrast, discriminatory explanations influenced how much they were blamed (Figure 5B; RQ1). This consideration had the largest effect on blame judgments of the AI system's developer and a significant but smaller effect on reactive attitudes towards its user. Perceived fairness did not impact judgments of the AI system; its effect size is close to zero across all studies.

Harmfulness influenced judgments of actors at different levels (Figure 5C; RQ2). It had the largest effect on the blame directed at the AI system's user, while its effect on judgments of the machine and its developer was smaller. Finally, the main factor at play in how people blame the AI system can be captured by people's explanations of their judgments (Figure 5D; RQ3). Those who did *not* mention the developer or user when blaming the AI system directed blame towards the machine, mitigating their reactive attitudes towards human stakeholders. This finding led to the hypothesis that those who perceive the AI system's behavior as constrained by its programming blame the machine less. However, Study 3 was not successful in manipulating perceived autonomy of the system. Considering the large effect sizes of people's justifications presented in Figure 5D, we used a qualitative assessment of participants' explanations to understand which specific factors are associated with how people blame AI systems.

5 参与者理由的定性分析

We inspected participants' explanations manually; two authors grouped each response into several categories. We focused on people's justifications for their blame judgments of AI systems. The procedure was divided into three main steps. First, the authors coded a random sample of 5% of responses to identify a series of categories; this initial procedure achieved high inter-coder reliability (Krippendorff's α = 0.879). Second, the authors coded an additional random sample of 10% of responses to inspect whether the codes were sufficient and complete. There were no changes to the code list, and this step achieved high reliability (α = 0.868). Finally, the remaining responses were divided in half and coded separately. Disagreements and uncertain cases were discussed after independent coding for agreement. Responses were allowed to be coded as more than one category. The coding procedure identified four categories:

- 机器：通过强调机器在决策过程中的作用（例如，"AI 作出了判断"）或明确表示"系统应对此负责"，来论证将责任归咎于 AI 系统的合理性。
- 并非机器：谴责将责任归咎于 AI 系统这一想法的回应，例如"不能把责任归咎于软件"或"不能让它承担责任"。

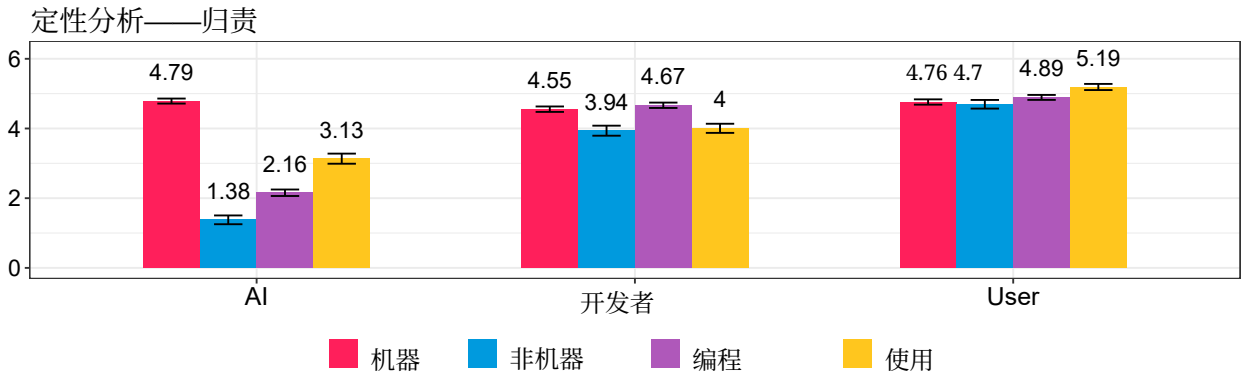

图 6：在所有研究中，根据参与者对其 AI 系统责备判断的辩解，对 AI 系统、其开发者和用户的责备判断。参与者的解释被定性编码为四类（见第 5 节）。标准误以误差条表示。

- 编程：强调开发者在这一过程中所起作用的回答（例如，“人工智能是由人类开发的”），或强调人工智能系统的行为受其编程约束的回答（例如，“该系统只是根据其编程方式行事”）。这些参与者强调了人工智能系统编程的某些方面。
- 使用情况：强调应有人监督 AI 系统的回答（例如，“仍应由人类审核这些决定”），或指出用户应承担责任的回答（例如，“操作人员负有责任”）。总体而言，这些回答侧重于 AI 系统的使用方式。

在本分析中，剔除了那些作出无法理解的回答，以及因认为该决定是正确的（例如“这是正确的做法”）而没有责怪任何人，或没有责怪特定主体（例如“等待时间太长”）的参与者——共160名参与者（13.87%）。

我们将参与者对AI系统、其开发者和用户的责备判断（采用 7 点量表）回归到四个虚拟变量上；这些变量根据参与者辩解的定性编码确定，符合条件时取值为 1（见表1）。我们还纳入了可解释性和危害性处理的虚拟变量（以及二者的交互项），以控制其影响（完整模型见附录中的表 14）。研究作为固定效应纳入模型。

Figure 6 presents mean values of the responses coded as each category. People's stance towards AI systems and the possibility of blaming them played a significant role in how they reacted to algorithmic harm. Those who believe one can blame machines (coded as Machine) directed their reactive attitudes at the AI system at a similar level to its designer and user. In contrast, a smaller but significant number of participants (Not Machine) denounced this idea and chose to attribute almost no blame to the machine. Blame towards the AI system was also associated with how people viewed other actors' roles in the decision-making process. Interestingly, harmfulness considerations, which were significant in our previous analyses, did not impact judgments of the AI system anymore (see Table 14 in the Appendix). We see that the R^2 of the model predicting blame towards AI is much higher than those of the regressions modeling blame judgments of developers and users,

	AI (1)	开发者 (2)	User (3)
机器（N = 417）	1.765***	0.328*	−0.040
	(0.158)	(0.154)	(0.140)
非机器 (N = 174)	−1.364***	−0.292†	0.021
	(0.168)	(0.163)	(0.149)
编程 (N = 443)	−1.019***	0.514***	0.125
	(0.141)	(0.137)	(0.125)
使用情况 (N = 194)	0.098	−0.364*	0.446***
	(0.149)	(0.145)	(0.132)
截距	3.135***	3.773***	4.088***
	(0.219)	(0.213)	(0.194)
调整后 R^2	0.438	0.105	0.077
观测值	993	993	993

Table 1: Regression analysis of blame judgments of the AI system, its developer, and user as a function of participants' justification of their blame judgments towards the AI system. We report the number of responses coded as each category in the first column. Standard errors are shown inside parentheses. We present the complete results in Table 14 in the Appendix. $^{\dagger}p$ <.1, $^{*}p$ <.05, $^{**}p$ <.01, $^{***}p$ <.001.

这表明，人们对“机器应否受责备”的立场，确实是影响人们是否以及如何责备 AI 系统的最重要因素。将这些因素从回归模型中剔除后，R^2 系数降至 0.032。

对用户的判断符合预期：那些强调用户在算法决策中所起作用的人（编码为 Usage），对用户的责备程度更高。同样，对 AI 开发者的责备判断受到人们所认为的开发者对 AI 系统决策的控制程度（Programming）的影响。公平性和伤害性考量仍然是预测对这两类行为者责备程度的重要因素。

然而，我们还观察到，强调人工智能系统能动性的人（编码为“机器”）更倾向于责怪开发者，而谴责对机器采取反应性态度的参与者（“非机器”）则较少责怪开发者。尽管样本很小。

degree, reactive attitudes towards the developer seem to track moral judgments of the AI system. It may well be that some participants equate the AI system and its developer (and not the user) when making moral judgments. This perception is often shared by scholars who denounce reactive attitudes towards AI; they often argue that blaming a machine would be an indirect way of blaming the developer with the caveat that the human stakeholder may be able to ultimately escape blame [18]. In contrast, blaming the AI system did not shift blame away from the developer in our studies.

Another possible explanation is that people's judgments are related to previous findings concerning blame towards supervisors for their subordinates' actions [43, 49]. This finding is also related to legal doctrines that hold superiors responsible for their subordinates' actions, e.g., the *respondeat superior* doctrine, which recent legal scholarship has proposed as a possible liability regime for AI [72, 98]. We cannot rule out any of these explanations. It is possible that both of them exist concurrently. Some people may equate the AI system to its developer, whereas others choose to blame the developer *alongside* its machine due to their relationship. Future work should delve deeper into how blame towards an AI system may correlate with judgments of its developer.

6 一般性讨论与启示

Our findings show that public reactions to algorithmic harm will not be restricted to only one actor but will be distributed across many entities. Both developers and users were blamed to a large extent in all scenarios, suggesting that both are targets of the public's reactive attitudes when AI systems cause harm. This result reinforces the "problem of many hands" [103], which posits the difficulty of pinpointing who is morally responsible when many actors are involved in an activity. People's reactive attitudes towards algorithmic harm seem to be distributed across various actors, including the machine itself, as a form of joint responsibility [50].

People's distribution of blame across a wide range of actors poses challenges to policymaking. Researchers have argued that the responsibility issues posed by autonomous machines should be viewed as political questions by weighing the conflicting views of a given community [91]. Disregarding the public's opinion when regulating AI systems could create a "law in the books," leading to a legal system that becomes unfamiliar to the people it aims to regulate [16]. Previous work has found that laypeople's reactions to algorithmic harm may conflict with existing liability models [69], highlighting the importance of understanding how public opinion may become a barrier to the successful regulation of AI systems. Adopting liability models that hold specific actors liable may conflict with people's expectations identified by our research. A possible starting point could be joint and several liability models, under which several parties can be held jointly liable in accordance to their respective obligations [107].

6.1 设计启示

公平性考量在对人工智能系统设计者的归责判断中发挥了重要作用，而感知到的伤害性在对用户的归责中则更为显著。尽管这两个概念可能相关，但它们之间的相互作用总体而言，这并未影响参与者对设计者或用户的判断。这一发现展现了不同的道德考量可能以独特的方式影响人们对部署和开发 AI 系统者的责备。

One crucial component of blame judgments is the extent to which an actor has control over an outcome [74]. Explainability can be viewed under the developers' control, as they can decide when and how explainability will be implemented (see more below). In contrast, users have more control over how AI systems will be deployed and thus how harmful their decisions may end up being. Hence, perceived control could prove to be a mediator of people's responses. This hypothesis is consistent with our results, which indicate that fairness considerations play a major role in blame judgments of the AI system's designer, whereas perceived harmfulness is more significant in blame directed at the user.

Design decisions concerning how much control users and developers have over AI systems could thus impact laypeople's reactions when things go awry. Human-AI collaborations, for instance, could shift blame towards users, whereas autonomous systems could highlight the responsibility of designers. This possibility is consistent with prior work investigating perceptions of responsibility for autonomous vehicles' crashes [68], which has found that users are deemed more responsible when cars are not completely autonomous, whereas developers receive more blame in cases of complete autonomy. Hence, deciding how much control users should have over consequential AI systems has not only the potential to impact their deployment but also who will be deemed accountable for any potential harm.

Developers and users were attributed higher levels of forward-looking responsibilities (i.e., obligations) than AI systems, showing that people expect human stakeholders to *take* responsibility for their systems. Even when people blame AI systems for their harmful actions, they still expect human stakeholders to ensure that machines are safe and make correct decisions. Other scholars have argued for developers to proactively take responsibility in the context of high-risk AI systems [55, 104]. These results highlight the public demand for developers and users to take on forward-looking responsibilities, meaning that people expect developers and users to take proactive actions to prevent potential harms.

6.1.1 Explainability and Accountability. **One of explainability's primary goals is to ensure that humans maintain meaningful human control over algorithms and thus remain responsible [81, 89]. In contrast, our results suggest that explainability alone does not significantly influence whom people hold responsible for algorithmic harm.** Although discriminatory explanations increased blame directed at the developer and user, an AI system solely explaining its decision did not modify how any of the actors was blamed, i.e., their perceived responsibility. Although explainability may be required for holding human actors legally responsible, such a consideration does not seem to influence public reactions. It is important to note that we focused on one particular type of explanation, i.e., simple natural language explanations. Future work can explore whether our results can be extended to different methods of explaining algorithmic decisions.

可解释性赋予开发者一种新的权力，使他们能够控制哪些信息会被呈现出来给。

those subjected to algorithmic decision-making [10]. This power relation raises the question of whether designers could take advantage of their privileged position to shift perceived responsibility (and blame) away from themselves and towards other stakeholders [71]. For instance, developers could implement explanations that obscure the discriminatory nature of an algorithmic decision, mitigating the amount of blame they would normally receive and thus impacting to whom laypeople turn for answers (i.e., hold accountable) [26]. Fortunately, our results suggest these design decisions may have to be more intentional since the mere presence of explanations did not impact perceived responsibility.

需要补充说明的是，参与者希望开发者承担面向未来的责任，这表明人们期待开发者意识到自身掌握的权力，并且不滥用这种权力。因此，应当促进开发者之间的问责，使他们意识到 XAI 系统开发过程中涉及的权力关系以及公众的期望。

Other factors should also be taken into consideration when developing explainable AI systems. Recent work has proposed a series of interventions to ensure that users do not over-rely on explainable machines [19]. Scholars have also observed that technical knowledge about AI affects how people interpret explanations [34]. These factors may also influence how people react to algorithmic harm. For instance, users that over-rely on explainable machines may blame machines more, whereas people with a technical understanding of AI may direct their reactive attitudes towards other actors.

6.2 算法歧视

In contrast to its developer and user, an AI system was not blamed more when it justified its decision with a discriminatory explanation. This surprising result was found for both egregious and illegal explanations based on gender and lawful but discriminatory justifications relying on race. This finding highlights how people react to algorithmic harm differently than human-caused harm. Fairness is a crucial dimension of human moral reasoning and should increase the extent to which human actors are blamed [45]; we even observed the expected result concerning the AI system's developer and user. A possible explanation is that people attribute less discriminatory motivations to machines because they are perceived to lack the mental capacities necessary for holding prejudices, as suggested by Bigman et al. [12], leading to lower levels of blame.

Our research suggests that people may underreact to algorithmic discrimination compared to their reactions to discrimination by humans. Previous work has found that people feel less outraged when algorithms discriminate relative to humans and are more likely to endorse stereotypes after unfair decisions are made by an algorithm [12]. Similarly, a prior study suggests that people are more tolerant to biases and harms caused by AI because of techno-optimistic narratives [57]. Taking blame as a signal of commitment to a shared set of norms [95], our findings suggest that people do not react to algorithmic discrimination by committing to socially accepted fairness norms. In other words, people's reactive attitudes towards the AI system do not necessarily denounce algorithmic discrimination. Scholars have raised worries that "bad machines 败坏良好道德" [61]。同样地，我们发现，算法歧视可能会助长有害刻板印象的正常化。未来的研究可以进一步探讨这一假设，例如调查参与者自我报告的反应性态度是否能够被激发，并转化为真实的态度变化。

6.3 对人工智能系统的道德判断

This research showed that AI systems are often recognized as blameworthy agents alongside their designers and users. In line with previous work (e.g., [42, 58, 70, 76]), people blamed AI systems for the harm they caused. Our studies highlight how people employ different moral considerations when judging machines and humans. The degree of blame attributed to human (and collective) agents depended on the properties of the decision—its fairness and harmfulness. For machines, this was not the case. **People's reactive attitudes towards AI systems were most strongly influenced by whether they believed blaming machines would be a suitable response to algorithmic harm.**

当人们相信机器可以被归咎时，他们对人工智能系统的指责程度与对其他人类行为者的指责一样高。相比之下，那些否定这种可能性的人往往不会将其反应性态度指向这些系统。对机器的道德判断首先要求人们评估：将责备作为对算法伤害的回应是否恰当。我们确定了三种主要可能性：

（1）人们接受责备机器的观点，并像责备人类和集体行动者一样责备机器；（2）人们谴责这种可能性，不将其反应性态度指向机器；（3）或者，人们将人工智能系统视为其开发者和用户的下属，并对其进行适度责备。

How these considerations may interact or add up with other moral factors, e.g., fairness and harmfulness, is an important line of work to understand how the general public reacts to algorithmic harm. Our findings are consistent with previous work showing that one's perceptions of AI impacts how they view and use such systems [15, 35–37]. Lay perceptions of algorithms not only have the potential to impact their deployment but also people's reactions to circumstances where they cause harm to individuals.

Our research does not argue for any of the possibilities above. Instead, our studies elucidate how people react to algorithmic harm in distinctive ways so that policymakers and developers can be aware of folk attitudes, allowing AI systems to be developed and deployed safely. A common concern is that developers and users will ultimately escape blame if people direct their reactive attitudes towards the AI system. Our findings do not support these concerns; instead, they show that those who decided to blame AI systems also blamed their developers more. Future research could delve deeper into whether our results are replicated in different environments.

我们的研究结果凸显了人们对人工智能系统的认知如何影响他们对算法伤害的反应。那些将机器视为道德共同体中新型行动者的人，会接受机器应承担的责任。相反，那些将机器视为人类工具的人，则选择凸显参与算法决策的人类。未来关于如何引入人工智能系统的决策将……

社会，例如，是否应将其纳入应受责备的行为主体之列，将是人们对算法伤害作出反应时的主要影响因素。这些决策应考虑到，公众反应可能会对这些新技术的成功治理带来挑战。政治审议是缓解公众与社会及监管机构之间可能使人工智能治理陷入不稳定的任何冲突的重要工具 [91]。

7 结语

Our research did not introduce a particular definition of blame to participants. This approach provides good external validity as blame is often ill-defined in real-life circumstances, allowing participants to have different opinions about whether to blame machines. However, we could not observe whether different participants use the same notion of blame for all actors or whether their notions vary depending on who is being judged. Someone, for instance, may blame an AI system to signal their commitment to a shared set of norms [95] and, at the same time, blame developers to condemn immoral behavior [92]. Future work could explore how introducing specific definitions of blame impacts whether and to what extent humans and machines are blamed. Similarly, studies could expand on our qualitative efforts to investigate which conceptions of blame are being relied on when people indicate AI systems, developers, and users should be blamed.

Our research was restricted to algorithmic decision-making, a significant use case of existing AI systems. We hence do not generalize our results to other scenarios. For instance, scholars have investigated public expectations of self-driving cars [7], and recent work has inquired how laypeople react to crashes through the lens of blame [41]. Different research fronts are necessary to understand how people react to algorithmic harm. Although we have examined how people's reactive attitudes are influenced by a series of moral considerations, e.g., harmfulness and fairness, other factors may also impact the extent to which AI systems and human stakeholders are blamed, such as perceived intentionality, which is a crucial component of moral judgments [60, 74].

Another set of factors that may influence blame judgments are individuals' demographics. Prior research has found that people's demographics are correlated with their perceptions of algorithmic fairness. Pierson [86] found that women are less likely than men to support including gender as a feature in decision-making algorithms used in the education domain. Grgić-Hlača et al. [48] did not find the same gender effect in the context of bail decisions, but they did find a correlation between political leaning and perceptions of fairness. These findings suggest that demographics may correlate with blame judgments. In the context of our study, gender may be correlated with reactive attitudes towards gender-based discrimination, and race with reactions to race-based discrimination.

同样，算法决策的对象也可能影响普通人对算法伤害的反应。对于公平性的看法，会因歧视性决策针对的是自己还是他人而有所不同 [99]。在责任归属方面，既有研究表明，与仅作为旁观者相比，当人们在经济博弈中自身利益受损时，他们更愿意惩罚相关主体 [39]。因此，未来研究可以探讨参与者是否会将更高的……

如果他们是算法歧视的受害者，对人工智能系统、开发者和用户的责备程度。

We have relied on previous work to demonstrate that people employ different moral considerations when judging machines and humans. Future work may experimentally compare moral judgments of both actors under circumstances similar to ours (e.g., see [51, 70]). We also measured people's reactive attitudes through vignette-based self-reported measures. Another possibility would be to conduct studies where participants interact with real AI systems and demonstrate their reactive attitudes through behavioral measures, e.g., whether or not they would use or cooperate with an AI system after failures (e.g., [30, 33, 52]). Finally, we studied how people react to AI systems taking on the role of decision-makers. As discussed in Section 2.5, several algorithms are deployed alongside humans and do not necessarily have the final say in decisions. Understanding whether our results are replicated in human-AI collaborations is a significant line of work to comprehend how the general public reacts to algorithmic harm.

Our study showed that people do not agree on whether to blame AI systems, reflecting the longstanding philosophical debate concerning the appropriateness of people's reactive attitudes towards machines. As argued by proponents of the property view of blame [101, 106], some people underscore that AI systems do not satisfy some preconditions for receiving blame and choose not to direct reactive attitudes towards them. In contrast, others view machines as actors embedded in social structures that call for blame when harmful decisions are made (e.g., the medical domain), which is consistent with the social view of blame [25, 100]. Philosophers discussing who (and what) can be blamed could consider studies like ours to revisit their assumptions and conclusions about what blame is and should be.

Our studies showcase how people react to algorithmic harm in unexpected ways. Factors that influence attribution of blame to human actors, such as fairness, did not influence people's reactive attitudes towards AI systems. Instead, our findings show that moral judgments of AI systems are determined by people's stance towards the possibility of blaming machines for algorithmic harm. Deciding whether and how to include machines into the social and moral spheres will shape how the general public reacts to them and their actions. Most importantly, this undertaking should account for how the public response may clash with the governance necessary for deploying AI systems safely in the real world.

致谢

本研究得到基础科学研究院（IBS- R029-C2）和韩国国家研究基金会（RS-2022- 00165347）的支持。

参考文献

[1] Ameena T Ahmed, Selina A Mohammed, David R Williams, et al. 2007. Racial discrimination & health: Pathways & evidence. *Indian Journal of Medical Research* 126, 4 (2007), 318.

[2] Salman Ahmed, Cameron T Nutt, Nwamaka D Eneanya, Peter P Reese, Karthik Sivashanker, Michelle Morse, Thomas Sequist, and Mallika L Mendu. 2021. Examining the potential impact of race multiplier utilization in estimated glomerular filtration rate calculation on African-American care outcomes. *Journal of General Internal Medicine* 36, 2 (2021), 464–471.

[3] Athanasios Andreou, Giridhari Venkatadri, Oana Goga, Krishna Gummadi, Patrick Loiseau, and Alan Mislove. 2018. Investigating ad transparency mechanisms in social media: A case study of Facebook's explanations. In *proc. of the Network and Distributed System Security Symposium*. 1–15.

[4] 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. ProPublica. https://tinyurl.com/5t3apr69.

[5] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[6] Edmond Awad, Sohan Dsouza, Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2020. Crowdsourcing moral machines. *Commun. ACM* 63, 3 (2020), 48–55.

[7] 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.

[8] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *proc. of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[9] Gagan Bansal, Tongshuang Wu, Joyce Zhu, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S Weld. 2020. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. *arXiv preprint arXiv:2006.14779* (2020).

[10] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[11] Adrien Bibal, Michael Lognoul, Alexandre de Streel, and Benoît Frénay. 2020. Legal requirements on explainability in machine learning. *Artificial Intelligence and Law* (2020), 1–21.

[12] Yochanan Bigman, Kurt Gray, Adam Waytz, Mads Arnestad, and Desman Wilson. [n.d.]. Algorithmic discrimination causes less moral outrage than human discrimination. ([n. d.]).

[13] Yochanan E Bigman, Adam Waytz, Ron Alterovitz, and Kurt Gray. 2019. Holding robots responsible: The elements of machine morality. *Trends in cognitive sciences* 23, 5 (2019), 365–368.

[14] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2020. *The moral psychology of AI and the ethical opt-out problem.* Oxford University Press.

[15] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *proc. of the ACM Conference on Human Factors in Computing Systems*. 1–12.

[16] Bartosz Brożek and Bartosz Janik. 2019. Can artificial intelligences be moral agents? *New Ideas in Psychology* 54 (2019), 101–106.

[17] Joanna J Bryson. 2010. Robots should be slaves. *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues* 8 (2010), 63–74.

[18] 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.

[19] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *proc. of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[20] Noah Castelo, Maarten W Bos, and Donald R Lehmann. 2019. Task-dependent algorithm aversion. *Journal of Marketing Research* 56, 5 (2019), 809–825.

[21] Stephen Cave, Claire Craig, Kanta Dihal, Sarah Dillon, Jessica Montgomery, Beth Singler, and Lindsay Taylor. 2018. Portrayals and perceptions of AI and why they matter. (2018).

[22] Stephen Cave and Kanta Dihal. 2019. Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence* 1, 2 (2019), 74–78.

[23] Marc Champagne. 2021. The Mandatory Ontology of Robot Responsibility. *Cambridge Quarterly of Healthcare Ethics* 30, 3 (2021), 448–454.

[24] Karen W Clarke, David Gray, Nicola A Keating, and John R Hampton. 1994. Do women with acute myocardial infarction receive the same treatment as men? *BMJ* 309, 6954 (1994), 563–566.

[25] 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.

[26] A Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. 2022. Accountability in an algorithmic society: relationality, responsibility, and robustness in machine learning. In *proc. of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. 864–876.

[27] John Danaher. 2016. Robots, law and the retribution gap. *Ethics and Information Technology* 18, 4 (2016), 299–309.

[28] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://tinyurl.com/y64598bh.

[29] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proceedings on Privacy Enhancing Technologies* 1 (2015), 92–112.

[30] Celso M de Melo, Stacy Marsella, and Jonathan Gratch. 2019. Human cooperation when acting through autonomous machines. *PNAS* 116, 9 (2019), 3482–3487.

[31] Filippo Santoni de Sio and Giulio Mecacci. 2021. Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology* (2021), 1–28.

[32] Berkeley J Dietvorst and Soaham Bharti. 2020. People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological Science* 31, 10 (2020), 1302–1314.

[33] Elias Fernández Domingos, Inês Terrucha, Rémi Suchon, Jelena Grujić, Juan C Burguillo, Francisco C Santos, and Tom Lenaerts. 2021. Delegation to autonomous agents promotes cooperation in collective-risk dilemmas. *arXiv preprint arXiv:2103.07710* (2021).

[34] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O Riedl, et al. 2021. The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations. *arXiv preprint arXiv:2107.13509* (2021).

[35] Motahhare Eslami, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016. First I" like" it, then I hide it: Folk Theories of Social Feeds. In *proc. of the ACM Conference on Human Factors in Computing Systems*. 2371–2382.

[36] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. " I always assumed that I wasn't really that close to [her]" Reasoning about Invisible Algorithms in News Feeds. In *proc. of the ACM Conference on Human Factors in Computing Systems*. 153–162.

[37] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User attitudes towards algorithmic opacity and transparency in online reviewing platforms. In *proc. of the ACM Conference on Human Factors in Computing Systems*. 1–14.

[38] 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.

[39] Ernst Fehr and Urs Fischbacher. 2004. Third-party punishment and social norms. *Evolution and Human Behavior* 25, 2 (2004), 63–87.

[40] Nic Fleming. 2018. How artificial intelligence is changing drug discovery. *Nature* 557, 7706 (2018), S55–S55.

[41] Matija Franklin, Edmond Awad, and David Lagnado. 2021. Blaming automated vehicles in difficult situations. *Iscience* 24, 4 (2021), 102252.

[42] Caleb Furlough, Thomas Stokes, and Douglas J Gillan. 2021. Attributing blame to robots: I. The influence of robot autonomy. *Human factors* 63, 4 (2021), 592–602.

[43] Donald E Gibson and Scott J Schroeder. 2003. Who ought to be blamed? The effect of organizational roles on blame and credit attributions. *International Journal of Conflict Management* (2003).

[44] Jin X Goh, Judith A Hall, and Robert Rosenthal. 2016. Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass* 10, 10 (2016), 535–549.

[45] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*. Vol. 47. Elsevier, 55–130.

[46] Kurt Gray, Liane Young, and Adam Waytz. 2012. Mind perception is the essence of morality. *Psychological inquiry* 23, 2 (2012), 101–124.

[47] Nina Grgić-Hlača, Christoph Engel, and Krishna P Gummadi. 2019. Human decision making with machine assistance: An experiment on bailing and jailing. *proc. of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.

[48] Nina Grgić-Hlača, Gabriel Lima, Adrian Weller, and Elissa M. Redmiles. 2020. Dimensions of Diversity in Human Perceptions of Algorithmic Fairness. https://doi.org/10.48550/ARXIV.2005.00808

[49] V Lee Hamilton. 1986. Chains of Command: Responsibility Attribution in Hierarchies 1. *Journal of Applied Social Psychology* 16, 2 (1986), 118–138.

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

[51] César A Hidalgo, Diana Orghiain, Jordi Albo Canals, Filipa De Almeida, and Natalia Martin. 2021. *How humans judge machines.* MIT Press.

[52] Fatimah Ishowo-Oloko, Jean-François Bonnefon, Zakariyah Soroye, Jacob Crandall, Iyad Rahwan, and Talal Rahwan. 2019. Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence* 1, 11 (2019), 517–521.

[53] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[54] Deborah G Johnson. 2006. Computer systems: Moral entities but not moral agents. *Ethics and Information Technology* 8, 4 (2006), 195–204.

[55] Deborah G Johnson. 2015. Technology with no human responsibility? *Journal of Business Ethics* 127, 4 (2015), 707–715.

[56] Peter H Kahn Jr, Takayuki Kanda, Hiroshi Ishiguro, Brian T Gill, Jolina H Ruckert, Solace Shen, Heather E Gary, Aimee L Reichert, Nathan G Freier, and

Rachel L Severson. 2012. Do people hold a humanoid robot morally accountable for the harm it causes?. In *proc. of the ACM/IEEE International Conference on Human-Robot Interaction*. 33–40.
[57] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. " Because AI is 100% right and safe": User Attitudes and Sources of AI Authority in India. In *proc. of the CHI Conference on Human Factors in Computing Systems*. 1–18.
[58] Taemie Kim and Pamela Hinds. 2006. Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In *proc. of the IEEE International Symposium on Robot and Human Interactive Communication*. 80–85.
[59] Lauren Kirchner. 2020. Can Algorithms Violate Fair Housing Laws? The Markup. https://tinyurl.com/mr49a7yd.
[60] Joshua Knobe. 2003. Intentional action and side effects in ordinary language. *Analysis* 63, 3 (2003), 190–194.
[61] Nils Köbis, Jean-François Bonnefon, and Iyad Rahwan. 2021. Bad machines corrupt good morals. *Nature Human Behaviour* 5, 6 (2021), 679–685.
[62] Moritz Körber, Eva Baseler, and Klaus Bengler. 2018. Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied Ergonomics* 66 (2018), 18–31.
[63] Markus Langer, Tim Hunsicker, Tina Feldkamp, Cornelius J. König, and Nina Grgić-Hlača. 2021. "Look! It's a Computer Program! It's an Algorithm! It's AI!": Does Terminology Affect Human Perceptions and Evaluations of Intelligent Systems? arXiv:2108.11486 [cs.HC]
[64] Markus Langer, Cornelius J König, and Maria Papathanasiou. 2019. Highly automated job interviews: Acceptance under the influence of stakes. *International Journal of Selection and Assessment* 27, 3 (2019), 217–234.
[65] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (2021), 103473.
[66] Minha Lee, Peter Ruijten, Lily Frank, Yvonne de Kort, and Wijnand IJsselsteijn. 2021. People May Punish, But Not Blame Robots. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11.
[67] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 205395171875668 4.
[68] 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. *SAE Technical paper* 10 (2016), 2016–01.
[69] Gabriel Lima, Meeyoung Cha, Chihyung Jeon, and Kyung Sin Park. 2021. The Conflict Between People's Urge to Punish AI and Legal Systems. *Frontiers in Robotics and AI* 8 (2021), 339. https://doi.org/10.3389/frobt.2021.756242
[70] 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 *proc. of the CHI Conference on Human Factors in Computing Systems*. 1–17.
[71] Gabriel Lima, Nina Grgić-Hlača, Jin Keun Jeong, and Meeyoung Cha. 2022. The Conflict Between Explainable and Accountable Decision-Making Algorithms. In *proc. of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.
[72] Anat Lior. 2019. AI entities as AI agents: Artificial intelligence liability and the AI respondeat superior analogy. *Mitchell Hamline L. Rev.* 46 (2019), 1043.
[73] Peng Liu and Yong Du. 2021. Blame Attribution Asymmetry in Human–Automation Cooperation. *Risk Analysis* (2021).
[74] Bertram F Malle, Steve Guglielmo, and Andrew E Monroe. 2014. A theory of blame. *Psychological Inquiry* 25, 2 (2014), 147–186.
[75] Bertram F Malle, Stuti Thapa Magar, and Matthias Scheutz. 2019. AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. In *Robotics and well-being*. Springer, 111–133.
[76] 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 *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 117–124.
[77] Andreas Matthias. 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology* 6, 3 (2004), 175–183.
[78] Blakeley B McShane and Ulf Böckenholt. 2017. Single-paper meta-analysis: Benefits for study summary, theory testing, and replicability. *Journal of Consumer Research* 43, 6 (2017), 1048–1063.
[79] Laxmi S Mehta, Theresa M Beckie, Holli A DeVon, Cindy L Grines, Harlan M Krumholz, Michelle N Johnson, Kathryn J Lindley, Viola Vaccarino, Tracy Y Wang, Karol E Watson, et al. 2016. Acute myocardial infarction in women: a scientific statement from the American Heart Association. *Circulation* 133, 9 (2016), 916–947.
[80] Danaë Metaxa, Michelle A Gan, Su Goh, Jeff Hancock, and James A Landay. 2021. An Image of Society: Gender and Racial Representation and Impact in Image Search Results for Occupations. *proc. of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.
[81] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.
[82] Keith C Norris, Nwamaka D Eneanya, and L Ebony Boulware. 2021. Removal of race from estimates of kidney function: first, do no harm. *JAMA* 325, 2 (2021), 135–137.
[83] 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.
[84] Institute of Medicine (US). Committee on Women's Health Research. 2010. *Women's health research: Progress, pitfalls, and promise*. National Academies Press.
[85] Stefan Palan and Christian Schitter. 2018. Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27.
[86] Emma Pierson. 2017. Gender differences in beliefs about algorithmic fairness. (2017). arXiv:1712.09124
[87] Bryan Pietsch. 2021. 2 Killed in Driverless Tesla Car Crash, Officials Say. New York Times. https://tinyurl.com/3r85htt4.
[88] Angelisa C Plane, Elissa M Redmiles, Michelle L Mazurek, and Michael Carl Tschantz. 2017. Exploring user perceptions of discrimination in online targeted advertising. In *proc. of the USENIX Security Symposium*. 935–951.
[89] Scott Robbins. 2019. A misdirected principle with a catch: explicability for AI. *Minds and Machines* 29, 4 (2019), 495–514.
[90] Simon Romero. 2018. Wielding Rocks and Knives, Arizonans Attack Self-Driving Cars. New York Times. https://www.nytimes.com/2018/12/31/us/waymo-self-driving-cars-arizona-attacks.html.
[91] Henrik Skaug Sætra. 2021. Confounding complexity of machine action: a hobbesian account of machine responsibility. *International Journal of Technoethics (IJT)* 12, 1 (2021), 87–100.
[92] Thomas Michael Scanlon. 2013. Interpreting blame. *Blame. Its nature and norms* (2013), 84–99.
[93] Matthias Scheutz and Bertram F Malle. 2020. May machines take lives to save lives? Human perceptions of autonomous robots (with the capacity to kill). *Lethal autonomous weapons: Re-examining the law & ethics of robotic warfare* (2020).
[94] Guido Schwarzer et al. 2007. meta: An R package for meta-analysis. *R news* 7, 3 (2007), 40–45.
[95] David Shoemaker and Manuel Vargas. 2021. Moral torch fishing: A signaling theory of blame. *Noûs* 55, 3 (2021), 581–602.
[96] 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.
[97] Peter Strawson. 2018. *Freedom and Resentment*. Cornell University Press.
[98] Hannah R Sullivan and Scott J Schweikart. 2019. Are current tort liability doctrines adequate for addressing injury caused by AI? *AMA Journal of Ethics* 21, 2 (2019), 160–166.
[99] Leigh Thompson and George Loewenstein. 1992. Egocentric interpretations of fairness and interpersonal conflict. *Organizational Behavior and Human Decision Processes* 51, 2 (1992), 176–197.
[100] Daniel W Tigard. 2021. Artificial moral responsibility: How we can and cannot hold machines responsible. *Cambridge Quarterly of Healthcare Ethics* 30, 3 (2021), 435–447.
[101] Steve Torrance. 2008. Ethics and consciousness in artificial agents. *Ai & Society* 22, 4 (2008), 495–521.
[102] Ibo Van de Poel. 2011. The relation between forward-looking and backward-looking responsibility. In *Moral Responsibility*. Springer, 37–52.
[103] Ibo Van de Poel. 2015. *Moral responsibility*. Routledge.
[104] Aimee van Wynsberghe. 2021. Responsible Robotics and Responsibility Attribution.
[105] Michael Veale and Frederik Zuiderveen Borgesius. 2021. Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International* 22, 4 (2021), 97–112.
[106] Carissa Véliz. 2021. Moral zombies: why algorithms are not moral agents. *AI & SOCIETY* (2021), 1–11.
[107] David C Vladeck. 2014. Machines without principals: liability rules and artificial intelligence. *Wash. L. Rev.* 89 (2014), 117.
[108] Daisuke Wakabayashi. 2018. Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam. New York Times. https://tinyurl.com/58wsubvd.
[109] Julie Weed. 2021. Résumé-Writing Tips to Help You Get Past the A.I. Gatekeepers. New York Times. https://tinyurl.com/yc2hz9tp.
[110] Deborah R Zucker, John L Griffith, Joni R Beshansky, and Harry P Selker. 1997. Presentations of acute myocardial infarction in men and women. *Journal of General Internal Medicine* 12, 2 (1997), 79–87.

A 材料与测量

A.1 研究 1

在同意研究条款并阅读一段简短的研究介绍后，参与者阅读了以下情境描述：

> Systemy 是一家开发软件的本地科技公司。该公司正在扩张，想要招聘新的软件开发人员。Systemy 正在使用人工智能（AI）系统来决定招聘哪些人。
>
> Taylor 是一名刚大学毕业的初级女性软件开发人员。她决定申请 Sys- temy 的软件开发人员职位。一周后，她收到了以下录用决定邮件：
>
> *[Unfortunately, we are not able to offer you a position at the moment. Thank you for your interest in the position. / Unfortunately, we are not able to offer you a position at the moment. Our hiring AI system has decided this because you do not have the necessary experience. Thank you for your interest in the position. / Unfortunately, we are not able to offer you a position at the moment. Our hiring AI system has decided this because you are a woman. Thank you for your interest in the position.]*

参与者以被试间方式随机分配到三个不同情境之一。阅读情境后，参与者回答了以下问题：

(1) How much blame does the *AI system / AI system's developer / Systemy* deserve for the decision? → 7-point scale, anchored at 0 = No blame at all, 6 = Extreme blame.
(2) How responsible was the *AI system / AI system's developer / Systemy* for ensuring that it was the correct decision? → 7-point scale, anchored at 0 = Not responsible at all, 6 = Extremely responsible.
(3) Why? Please explain your choices in 1-2 sentences. → Open-ended question.

问题1-3针对每个行为主体分别呈现；行为主体的呈现顺序在参与者之间随机安排。随后，参与者回答了另一组问题：

(4) How harmful was the AI system's decision? → 7-point scale, anchored at 0 = Not harmful at all, 6 = Extremely harmful.
(5) To what extent did the AI system provide an explanation of its decision? → 7-point scale, anchored at 0 = Definitely not, 6 = Definitely yes.
(6) How fair was the AI system's decision? → 7-point scale, anchored at 0 = Not fair at all, 6 = Extremely fair.
(7) To what extent do you think the AI system's decision could apply to you? → 7-point scale, anchored at 0 = Definitely not, 6 = Definitely yes.
(8) How morally wrong was the AI system's decision? → 7-point scale, anchored at 0 = Not wrong at all, 6 = Extremely wrong.
(9) To what extent do you think the AI system could have made a different decision? → 7-point scale, anchored at 0 = Definitely not, 6 = Definitely yes.
(10) AI系统的决策有多大程度上是有意为之？ → 7点评分量表，端点为 0 =完全不是有意的， 6 = 极其有意。

问题4-7和8-10被分成两个不同的页面；这两个页面的顺序在参与者之间随机安排。

A.2 研究2

研究 2 向参与者呈现了以下情境：

> West Medical 是一家当地医院。West Medical 正在使用一套人工智能系统，评估前来急诊室（ER）的患者病情。
> 泰勒是一名女性，左臂不适并持续恶心三天后，决定前往西部医疗中心就诊。AI系统初步评估后，将泰勒归类为非危重患者，并要求她等待。
>
> *[/ 人工智能系统给出的决策理由如下：该患者被归类为非危重患者，因为其没有任何既往病症或合并症。/ 人工智能系统给出的决策理由如下：该患者被归类为非危重患者，因为该患者是女性。]*
> *[等待了八个小时后，泰勒接受了医生的检查，并拿到了缓解症状的药物处方。/ 等待了八个小时后，泰勒突发心脏病，在没有得到适当医疗救治的情况下死亡。]*

被分配到不可解释处理条件的参与者没有看到任何关于人工智能系统判定依据的信息。随后，参与者回答了与研究1相同的问题。

A.3 研究 3

研究3向参与者呈现了以下情境：

> West Medical 是一家地方医院。West Medical 正在使用一套 AI 系统，评估前往急诊室（ER）就诊的患者的病情。
> *[/ 该人工智能系统使用过去患者优先排序决策的数据开发而成。该系统学会了如何决定应优先安排哪位患者，而其开发过程中相关人员并未设定明确规则。]*
> 泰勒是 *[一名女性/ 一名非裔美国人]* 她在连续三天感到 *[左臂不适和恶心/ 感到疲倦、排尿频率降低且双腿肿胀]* 后，决定前往西部医疗中心。在接受 AI 系统的初步评估后，泰勒被归类为非危重患者，并被要求等待。
>
> *[/ AI 系统对其决策作出如下解释：该患者被归类为非危重患者，因为该患者没有任何既往病症或合并症。/ AI 系统对其决策作出如下解释：该*

患者被归类为非危重，因为该患者是女性/非裔美国人。*]*
*[*等待八小时后，泰勒接受了医生检查，并获得了缓解症状的药物处方。/ 等待八小时后，泰勒 *[*突发心脏病去世 / 因急性肾衰竭去世*]*，未能获得适当的医疗护理。*]*

Participants assigned to the non-explainable treatment condition were not shown any information concerning the AI system's justification. We employed the same design choice for the autonomy treatment. Participants read one of 2 (gender or race discrimination) x 2 (heart attack or kidney failure) vignettes in a between-subjects fashion. Participants then responded to the same questions as in Study 2. Additionally, they were asked the following questions after Q1-3 and before Q4-10.

(11) To what extent do you think the AI system's decision was under control of its developer? → 7-point scale, anchored at 0 = Definitely not, 6 = Definitely yes.
(12) To what extent do you think the AI system was programmed to make this decision? → 7-point scale, anchored at 0 = Definitely not, 6 = Definitely yes.
(13) To what extent do you think the AI system's decision was explicitly programmed by its developer? → 7-point scale, anchored at 0 = Definitely not, 6 = Definitely yes.

B 补充分析

变量	处理	N	AI		开发者		User	
			M	SD	M	SD	M	SD
可解释性	无解释	69	3.25	2.17	3.51	1.88	3.91	2.06
	解释	68	2.68	2.17	3.78	1.86	3.90	1.86
	歧视性解释	72	3.18	2.32	5.25	1.40	5.07	1.62
人们的辩解	未提及	118	3.41	2.22	4.07	1.88	4.20	1.86
	对开发者或用户的提及	91	2.56	2.16	4.36	1.88	4.44	2.01

表 2：研究 1 中，不同处理条件下对人工智能系统、其开发者及用户的责备判断均值（M）和标准差（SD）。图 2 以可视化方式呈现均值。

	差异。	*p*	95%置信区间
AI			
解释 – 无解释	-0.57	0.29	[-1.47,0.33]
歧视性解释 –无解释	-0.07	0.98	[-0.95,0.82]
歧视性解释 –解释	0.50	0.38	[-0.38,1.39]
开发者			
解释 – 无解释	0.27	0.62	[-0.42,0.97]
歧视性解释 –无解释	1.74	0.00	[1.06,2.43]
歧视性解释 –解释	1.47	0.00	[0.78,2.16]
User			
解释 – 无解释	-0.02	1.00	[-0.76,0.73]
歧视性解释 –无解释	1.16	0.00	[0.42,1.89]
歧视性解释 –解释	1.17	0.00	[0.43,1.91]

表3：研究1中可解释性处理条件之间责备判断的Tukey HSD事后检验。

变量	处理条件	N	AI		开发者		User	
			M	SD	M	SD	M	SD
可解释性	无解释	135	3.62	2.03	4.27	1.79	4.74	1.62
	有解释	136	3.04	2.17	3.96	1.74	4.54	1.64
	歧视性解释	135	3.71	2.31	4.96	1.42	5.06	1.24
有害程度	N/A	202	3.15	2.16	4.16	1.87	4.40	1.74
	有害	204	3.75	2.18	4.63	1.50	5.16	1.17
人们的理由	未提及	257	3.88	2.07	4.24	1.71	4.62	1.61
	提及开发者或用户	149	2.71	2.19	4.66	1.68	5.07	1.32

表 4：研究 2 中，不同实验条件下对 AI 系统、其开发者和用户的责备判断均值（M）和标准差（SD）。图 3 以可视化形式呈现了均值。

	差值	p	95% 置信区间
AI			
有解释 – 无解释	-0.58	0.06	[-1.17,0.01]
歧视性解释 –无解释	0.10	0.92	[-0.50,0.69]
歧视性解释 –解释	0.67	0.02	[0.08,1.27]
开发者			
解释 – 无解释	-0.31	0.27	[-0.78,0.16]
歧视性解释 –无解释	0.70	0.00	[0.23,1.17]
歧视性解释 –解释	1.01	0.00	[0.54,1.48]
User			
解释 – 无解释	-0.20	0.51	[-0.62,0.22]
歧视性解释 –无解释	0.32	0.18	[-0.10,0.74]
歧视性解释 –解释	0.52	0.01	[0.10,0.93]

表 5：Tukey 的 研究中可解释性处理条件间责备判断的 HSD 事后检验 **2.**

变量	处理条件	N	AI		开发者		User	
			M	SD	M	SD	M	SD
可解释性	无解释	196	3.05	1.98	3.88	1.69	4.65	1.47
	解释	175	3.19	2.23	4.02	1.81	4.69	1.68
	歧视性解释	167	3.17	2.27	4.88	1.63	4.93	1.55
有害性	N/A	274	2.74	2.08	4.02	1.84	4.35	1.78
	有害	264	3.54	2.16	4.47	1.64	5.17	1.17
人们的辩解	未提及	319	3.54	2.16	4.13	1.79	4.61	1.65
	提及开发者或用户	219	2.53	1.99	4.39	1.71	4.95	1.42

表 6：研究 3 中，根据处理条件，对 AI 系统、其开发者和用户的责备判断均值（M）和标准差（SD）。图 4 以可视化方式呈现了均值。

	差异。	p	95% 置信区间
AI			
说明 – 无说明	0.14	0.78	[-0.36,0.65]
歧视性解释 –无解释	0.13	0.83	[-0.38,0.64]
歧视性解释 –解释	-0.01	1.00	[-0.54,0.51]
开发者			
说明 – 无说明	0.14	0.70	[-0.27,0.55]
歧视性解释 –无解释	1.00	0.00	[0.58,1.41]
歧视性解释 –解释	0.86	0.00	[0.43,1.28]
User			
解释 – 无解释	0.04	0.97	[-0.33,0.40]
歧视性解释 –无解释	0.28	0.18	[-0.09,0.65]
歧视性解释 –解释	0.24	0.29	[-0.14,0.62]

表 7：研究 3 中不同可解释性处理条件下归责判断的 Tukey HSD 事后检验。

C 操纵检验分析

参数	平方_和	df	均方_	F	p	η_p^2
感知到的可解释性						
可解释性	214.17	2	107.09	24.89	0.00	0.19
残差	886.17	206	4.30			
感知到的公平性						
可解释性	241.22	2	120.61	50.16	0.00	0.33
残差	495.30	206	2.40			

表 8：研究 1 中感知可解释性和公平性的操纵检验分析。平均值请参见图 7。

	差异。	p	95%置信区间
感知到的可解释性			
解释 – 无解释	2.13	0.00	[1.29, 2.97]
歧视性解释 –无解释	2.17	0.00	[1.35, 3.00]
歧视性解释 –解释	0.04	0.99	[-0.78, 0.87]
感知公平性			
说明 – 无说明	0.60	0.06	[-0.03, 1.22]
歧视性解释 –无解释	-1.91	0.00	[-2.52,-1.29]
歧视性解释 –解释	-2.50	0.00	[-3.12,-1.89]

表 9：研究 1 中不同可解释性处理条件之间感知可解释性与公平性的 Tukey HSD 事后检验。

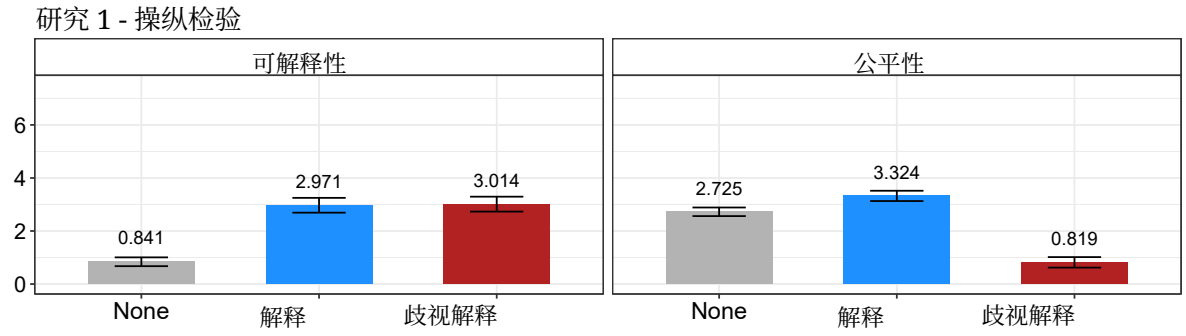

图 7：研究 1 中对感知可解释性和公平性的操纵检验分析。可解释的 AI 系统被认为更具可解释性，而不公平的解释则被认为更不公平。

参数	平方和	df	均方	F	p	η_p^2
感知到的可解释性						
可解释性	196.65	2	98.32	27.86	0.00	0.12
有害	0.17	1	0.17	0.05	0.83	0.00
可解释性：有害	6.37	2	3.19	0.90	0.41	0.00
残差	1411.46	400	3.53			
感知公平性						
可解释性	280.79	2	140.39	60.46	0.00	0.23
有害	87.16	1	87.16	37.54	0.00	0.09
可解释性：有害性	35.18	2	17.59	7.57	0.00	0.04
残差	928.86	400	2.32			
感知有害性						
可解释性	91.27	2	45.64	23.89	0.00	0.11
有害	491.23	1	491.23	257.21	0.00	0.39
可解释性：有害	49.11	2	24.55	12.86	0.00	0.06
残差	763.95	400	1.91			

表10：研究2中感知可解释性、公平性和危害性的操纵检验分析。平均值见图8。

	差异。	p	95%置信区间
感知可解释性			
解释 – 无解释	1.62	0.00	[1.08, 2.16]
歧视性解释 –无解释	1.27	0.00	[0.73, 1.80]
歧视性解释 –解释	-0.36	0.27	[-0.89, 0.18]
感知公平性			
解释 – 无解释	0.19	0.54	[-0.24, 0.63]
歧视性解释 –无解释	-1.66	0.00	[-2.10,-1.22]
歧视性解释 –解释	-1.85	0.00	[-2.29,-1.42]
感知到的危害性			
解释 – 无解释	0.24	0.34	[-0.16, 0.63]
歧视性解释 –无解释	1.10	0.00	[0.71, 1.50]
歧视性解释 –解释	0.87	0.00	[0.47, 1.26]

表 11：研究 2 中可解释性处理条件之间感知可解释性、公平性和危害性的 Tukey HSD 事后检验。

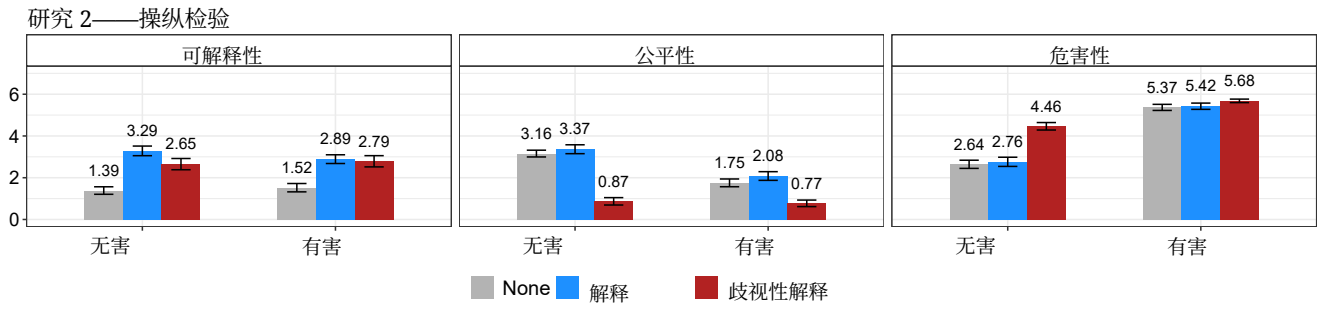

图 8：研究 2 中对感知可解释性、公平性和有害性的操纵检验分析。可解释人工智能系统被认为更具可解释性，不公平的解释被认为更不公平，而有害的决策则被认为危害更大。

参数	总和_平方	df	均_方	F	p	η_p^2
感知可解释性						
可解释性	226.31	2.00	113.16	31.58	0.00	0.11
有害	5.20	1.00	5.20	1.45	0.23	0.00
自主	0.00	1.00	0.00	0.00	0.99	0.00
情境短文	7.33	3.00	2.44	0.68	0.56	0.00
可解释性：有害	5.87	2.00	2.93	0.82	0.44	0.00
可解释性：自主	5.02	2.00	2.51	0.70	0.50	0.00
有害：自主性	1.86	1.00	1.86	0.52	0.47	0.00
可解释性：有害：自主性	10.73	2.00	5.36	1.50	0.22	0.01
残差	1873.75	523.00	3.58			
感知公平性						
可解释性	358.20	2.00	179.10	68.95	0.00	0.21
有害	171.40	1.00	171.40	65.98	0.00	0.11
自主的	2.98	1.00	2.98	1.15	0.28	0.00
情境短文	12.13	3.00	4.04	1.56	0.20	0.01
可解释性：有害	29.68	2.00	14.84	5.71	0.00	0.02
可解释性：自主	6.64	2.00	3.32	1.28	0.28	0.00
有害：自主的	0.02	1.00	0.02	0.01	0.93	0.00
可解释性:有害:自主	1.88	2.00	0.94	0.36	0.70	0.00
残差	1358.57	523.00	2.60			
感知到的危害性						
可解释性	105.04	2.00	52.52	26.20	0.00	0.09
有害	660.75	1.00	660.75	329.63	0.00	0.39
自主	2.78	1.00	2.78	1.38	0.24	0.00
情境描述	19.98	3.00	6.66	3.32	0.02	0.02
可解释性:有害	60.79	2.00	30.40	15.16	0.00	0.05
可解释性:自主	5.55	2.00	2.77	1.38	0.25	0.01
有害:自主	1.24	1.00	1.24	0.62	0.43	0.00
可解释性:有害:自主	2.36	2.00	1.18	0.59	0.56	0.00
残差	1048.37	523.00	2.00			
感知自主性						
可解释性	18.06	2.00	9.03	3.84	0.02	0.01
有害	7.13	1.00	7.13	3.03	0.08	0.01
自主	14.12	1.00	14.12	6.00	0.01	0.01
情境	13.32	3.00	4.44	1.89	0.13	0.01
可解释性：有害	1.62	2.00	0.81	0.34	0.71	0.00
可解释性：自主	8.03	2.00	4.02	1.71	0.18	0.01
有害:自主	0.52	1.00	0.52	0.22	0.64	0.00
可解释性：有害：自主性	4.03	2.00	2.01	0.86	0.43	0.00
残差	1229.68	523.00	2.35			

表 12：研究 3 中对感知可解释性、公平性、有害性和自主性的操纵检验分析。均值请参见图 9。

	差异。	p	95% 置信区间
感知到的可解释性			
解释 – 无解释	1.54	0.00	[1.08, 2.00]
歧视性解释 –无解释	1.09	0.00	[0.62, 1.56]
歧视性解释 –解释	-0.45	0.07	[-0.93, 0.03]
感知公平性			
解释 – 无解释	-0.04	0.98	[-0.43, 0.36]
歧视性解释 –无解释	-1.67	0.00	[-2.07,-1.27]
歧视性解释 –解释	-1.64	0.00	[-2.05,-1.23]
感知有害性			
解释 – 无解释	0.19	0.39	[-0.15, 0.54]
歧视性解释 –无解释	0.80	0.00	[0.45, 1.15]
歧视性解释 –解释	0.60	0.00	[0.24, 0.96]
感知自主性			
解释 – 无解释	0.32	0.11	[-0.05, 0.70]
歧视性解释 –无解释	0.47	0.01	[0.09, 0.85]
歧视性解释 –解释	0.15	0.65	[-0.24, 0.54]

表13：研究3中可解释性处理条件之间感知可解释性、公平性、有害性和自主性的Tukey HSD事后检验。

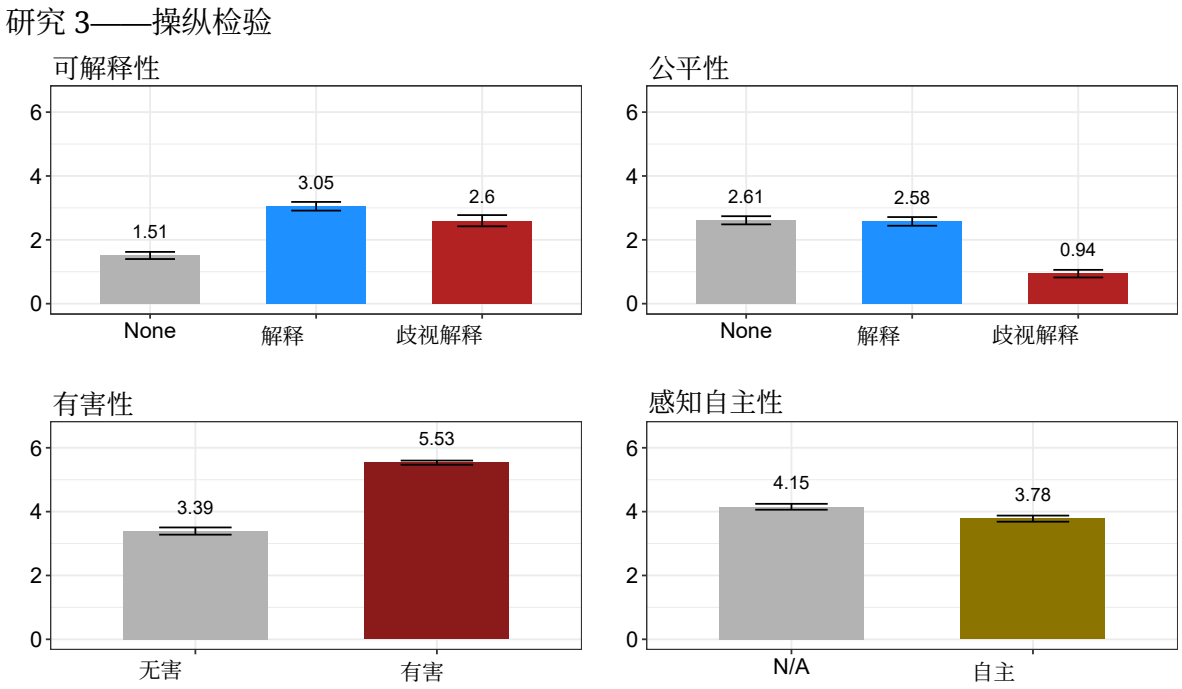

图9：研究3中对感知可解释性、公平性、有害性和自主性的操纵检验分析。可解释的人工智能系统被认为更具可解释性，不公平的解释被视为更加不公平，有害的决策被认为更具危害性。然而，我们的自主性处理未能达到足够强的效应量。我们在正文中呈现了不含该处理条件的结果。

	AI	开发者	User
	(1)	(2)	(3)
机器（N = 417）	1.765***	0.328*	−0.040
	(0.158)	(0.154)	(0.140)
非机器 (N = 174)	−1.364***	−0.292†	0.021
	(0.168)	(0.163)	(0.149)
编程（N = 443）	−1.019***	0.514***	0.125
	(0.141)	(0.137)	(0.125)
使用（N = 194）	0.098	−0.364*	0.446***
	(0.149)	(0.145)	(0.132)
有害（处理组）	0.303	0.400*	0.753***
	(0.191)	（0.185）	（0.169）
可解释性（处理组）	−0.205	−0.053	−0.010
	（0.176）	（0.171）	(0.156)
歧视（对待）	0.045	0.881***	0.560***
	(0.169)	(0.164)	(0.150)
研究 2	−0.068	0.029	0.164
	(0.163)	(0.159)	(0.145)
研究3	−0.258	−0.113	0.156
	(0.159)	(0.155)	(0.141)
有害性：可解释性	0.205	0.039	0.023
	(0.261)	(0.254)	(0.232)
有害：歧视	0.268	−0.121	−0.281
	(0.264)	(0.257)	(0.235)
截距	3.135***	3.773***	4.088***
	(0.219)	(0.213)	(0.194)
已调整 R^2	0.438	0.105	0.077
观察结果	993	993	993

Table 14: Regression analysis of blame judgments of the AI system, its developer, and user as a function of participants' explanation of their blame judgments towards the AI system. We report the number of responses coded as each category in the first column. Standard errors are shown inside parentheses. All responses in Study 1 were coded as not harmful, as their mean perceived harmfulness was closer to the perceived harmfulness of Study 2's and 3's not harmful conditions. $^{\dagger}p$ <.1, $^{*}p$ <.05, $^{**}p$ <.01, $^{***}p$ <.001.