Current Psychology
https://doi.org/10.1007/s12144-024-06658-2

Moral reasoning in a digital age: blaming artificial intelligence for incorrect high-risk decisions

Benedikt Leichtmann[1,3] · **Andreas Hinterreiter**[2] · **Christina Humer**[2] · **Alfio Ventura**[1,4] · **Marc Streit**[2] · **Martina Mara**[1]

Accepted: 3 September 2024

Abstract
The increasing involvement of Artificial Intelligence (AI) in moral decision situations raises the possibility of users attributing blame to AI-based systems for negative outcomes. In two experimental studies with a total of $N = 911$ participants, we explored the attribution of blame and underlying moral reasoning. Participants had to classify mushrooms in pictures as edible or poisonous with support of an AI-based app. Afterwards, participants read a fictitious scenario in which a misclassification due to an erroneous AI recommendation led to the poisoning of a person. In the first study, increased system transparency through explainable AI techniques reduced blaming of AI. A follow-up study showed that attribution of blame to each actor in the scenario depends on their perceived obligation and capacity to prevent such an event. Thus, blaming AI is indirectly associated with mind attribution and blaming oneself is associated with the capability to recognize a wrong classification. We discuss implications for future research on moral cognition in the context of human–AI interaction.

Keywords Moral psychology · Explainable artificial intelligence · Mind perception · Moral cognition · Mushroom picking game · Scapegoating

✉ Martina Mara
martina.mara@jku.at

Benedikt Leichtmann
benedikt.leichtmann@psy.lmu.de

Andreas Hinterreiter
andreas.hinterreiter@jku.at

Christina Humer
christina.humer@jku.at

Alfio Ventura
alfio.ventura@uni-due.de

Marc Streit
marc.streit@jku.at

[1] LIT Robopsychology Lab, Johannes Kepler University Linz, Altenberger Straße 69, Linz 4040, Austria

[2] Visual Data Science Lab, Institute of Computer Graphics, Johannes Kepler University Linz, Altenberger Straße 69, Linz 4040, Austria

[3] Department of Psychology, Ludwig-Maximilians-Universität München, Geschwister-Scholl-Platz 1, Munich 80539, Germany

[4] Research Center Trustworthy Data Science and Security, University of Duisburg-Essen, Bismarckstrasse 120, Duisburg 47057, Germany

Introduction: when humans blame AI

Artificial Intelligence (AI) is not only applied in harmless scenarios such as recommending new movies based on streaming histories. AI-based systems are also involved in moral situations with potentially severe consequences, such as car accidents in semi-automated driving (see recent publication in this journal, Copp et al., 2023). Here, the question arises: who do we blame for any harm?

Blame is a moral judgment and means "evaluating *agents* for their involvement in [...] norm-relevant events" (Malle et al., 2014, p.148). How people blame others may differ from legal and ethical standpoints on who *should* be held accountable (Komatsu et al., 2021). However, the empirical question of whom individuals blame and their *reasoning* for blaming is equally important, since our *subjective* moral judgments influence our actions. According to Greene (2015), the function of morality is to promote and sustain cooperation. Research has shown that blaming is negatively correlated with cooperation in games (Ketelaar & Tung Au, 2003; Nelissen et al., 2007) and positively correlated with the endorsement of punishment (Bastian et al., 2011;

Published online: 09 October 2024

Rothschild et al., 2012). Similarly, our blaming of different actors in human–technology interaction contexts might also lead to behavioral intentions (e.g., legal actions against the developer company).

Based on the *Path Model of Blame* by Malle et al. (2014), for humans to blame someone, they must first detect an event with a moral norm violation, and then determine that this event is caused by one or more agents. In situations of unintentional harm, persons evaluate each agent's obligation (= *should*) and capacity (= *could*) to prevent the event. Depending on the level of obligation and capacity, agents are then blamed to specific degrees.

With the increasing use of AI as decision support, blaming and moral reasoning are becoming important in the context of human–technology interaction (Shank & DeSanti, 2018; Sullivan & Fosso Wamba, 2022; Renier et al., 2021; Langer et al., 2021; Komatsu et al., 2021; Kim & Hinds, 2006; Hong et al., 2020; Bigman et al., 2019). One example scenario that creates moral situations is mushroom hunting. Previous work on human-AI interaction used the scenario of a person evaluating the edibility of mushrooms with the assistance of AI to study human decision-making, trust, and the effects of explainable AI (XAI) methods (i.e., an AI system explaining its classification decisions) (Leichtmann, Humer, et al., 2023; Leichtmann, Hinterreiter, et al., 2023; Humer et al., 2024). AI-supported mushroom picking bears the risk of false mushroom classification due to over-trust in system recommendations. Such mistakes can lead to serious health consequences if poisonous mushrooms are mistakenly consumed (Brandenburg & Ward, 2018; Schmutz et al., 2018; Cervellin et al., 2018). This poisoning of a human being creates a moral situation involving blaming processes.

In such a situation, multiple directly or indirectly involved actors could potentially be blamed. These entities include oneself, but also the AI-based system if it is identified as an actor within the causal chain. Two factors might influence the blaming of an AI:

First, AI can be perceived as an actor that is capable of making its own decisions. In doing so, users could ascribe a certain degree of mind to the machine. Research on mind perception of artificial agents such as robots, AI and other machines has rooted in the research tradition of theory of mind (see Gray et al., 2012; Gray et al., 2007; Gray and Wegner, 2009; Shank and DeSanti, 2018; Bigman et al., 2019; Waytz et al., 2010). According to this research line, human-like appearance or human-like behavior of machine systems could also lead to a greater attribution of human characteristics such as a mind of their own (Waytz et al., 2010). This involves the attribution of affect, the ability to interact independently with the environment, or the ability of mental and moral regulation (Malle, 2019).

Second, humans might "make" AI-based systems an actor in order to be able to "externalize blame for negative outcomes that would otherwise incriminate themselves or their group" (Rothschild et al., 2012, p.1149). This process of "scapegoating" is used to maintain one's own moral value (Rothschild et al., 2012).

Based on these two reasons for the attribution of blame to machines, XAI methods in which AI's decisions are explained to the end-user (Kim & Hinds, 2006) could have a special influence. In doing so, XAI methods reveal information about the causal history of a system classification (Miller, 2019), offering insights into how and why decisions are made. Various approaches, including text-based and visual explanations are employed to achieve this (Molnar, 2023; Barredo Arrieta et al., 2020; Guidotti et al., 2019). Examples for visual explanations are presenting a prototypical image from the training data (Jeyakumar et al., 2020) or highlighting the most decisive image regions (Selvaraju et al., 2017). XAI could make a system (*i*) appear more or less capable of preventing an outcome which makes the AI system seem more or less like an actor who can be blamed (i.e., if the system is perceived as an actor capable of preventing the event, it is blamed more), or (*ii*) influence people's perception that they themselves could have recognized an erroneous decision due to the explanation (e.g., if the explanation is not coherent and indicates unreliability), which in turn would make scapegoating more difficult. More precisely, one would no longer be able to blame AI alone, as soon as one becomes an actor oneself who would have been able to prevent the event by recognizing signs of misclassification in the incoherent explanation. This would then result in higher levels of self-blame and lower blame of AI. This mechanism could be attributable to XAI methods increasing the traceability of system classifications, meaning the ability for individuals to track and understand how classifications are made within a system, including the information processing involved (Schrills & Franke, 2023).

In this article, we describe the results of two empirical studies that used the scenario of mushroom hunting with assistance of an AI-based system in order to study blaming and moral reasoning in human-AI interaction after a negative outcome. The contribution of our work is three-fold:

1. Past studies show that explanations of AI decisions could change human perceptions and decision making in general (see e.g., Leichtmann, Humer, et al., 2023; Leichtmann, Hinterreiter, et al., 2023; Wischnewski et al., 2023; Yang et al., 2020). Thus, we tested whether such explanations of AI classifications also affect human blaming of AI and self-blame in a moral situation specifically (i.e., after an event of harm due to false classification)(study 1).

2. In a second step, we aimed to conceptually replicate this effect with other XAI methods to test its robustness and generalizability (study 2).
3. Additionally, we explored the associated moral cognition of blaming in human-AI interaction to understand underlying processes and reasons for variance in blaming. In particular, we explore the role of capability of the AI-based system and oneself to prevent the harm (study 2).

Research overview

We conducted two experiments to study blame attribution and moral reasoning in the context of human–AI interaction with the use case of mushroom picking. **Study 1** explored effects of XAI methods on blaming an AI-based system and oneself for negative consequences due to wrong decisions. **Study 2** aimed to (*i*) replicate this effect with three distinct XAI methods separately, (*ii*) explore underlying reasoning based on concepts associated with blaming a technological entity and scapegoating (e.g., mind perception or perceived possibility to prevent an outcome), and (*iii*) test how blame is spread across multiple actors directly and indirectly involved in the norm-relevant event (Gerstenberg & Lagnado, 2010; Shank & DeSanti, 2018).

Data collection occurred in two waves as part of a larger research project. Data unrelated to blaming, including results on effects of XAI methods on human decision-making or self-reported trust, have been discussed elsewhere (Leichtmann, Hinterreiter et al., 2023; Leichtmann, Humer, et al., 2023). Results presented here are original and have not been published before. All data and the analysis code can be found online at OSF (https://osf.io/375xu/).

Both studies complied with the tenets of the Declaration of Helsinki and adhered to ethical guidelines of the APA Code of Conduct. Informed consent was obtained from each participant prior to data collection.

Study 1: visual explanations reduce blame on AI

Methods

For our first analysis, we used data from an online between-subjects experiment with $N = 410$ participants (213 female, 193 male, 2 non-binary, 2 without gender specification; mean age $M = 44.58$ years, $SD = 15.29$). Participants were instructed to imagine a mushroom hunt with the goal to pick all edible mushrooms for a meal and leave inedible and poisonous ones (Leichtmann, Humer, et al., 2023). They had to classify mushrooms shown in 15 different photographs as edible or inedible/poisonous and indicated whether they would pick or leave them. For this task, participants were supported by an AI-based mushroom classification app (called "Forestly"). The experiment manipulated the AI-based system's explainability on two levels. One group of participants received the AI-based classification without further explanations ($n = 208$). The other group received a combination of two visual explanation strategies ($n = 202$): (*i*) the attribution-based technique GradCAM (Selvaraju et al., 2017) that highlights regions of an image important for a model's decision, and (*ii*) an adapted version of the example-based technique ExMatchina (Jeyakumar et al., 2020) that picks out specific data items as examples to be shown to users. Example images of the two interface variants are shown in Fig. 1.

After completing the mushroom-picking task and questionnaires, participants read a short vignette:

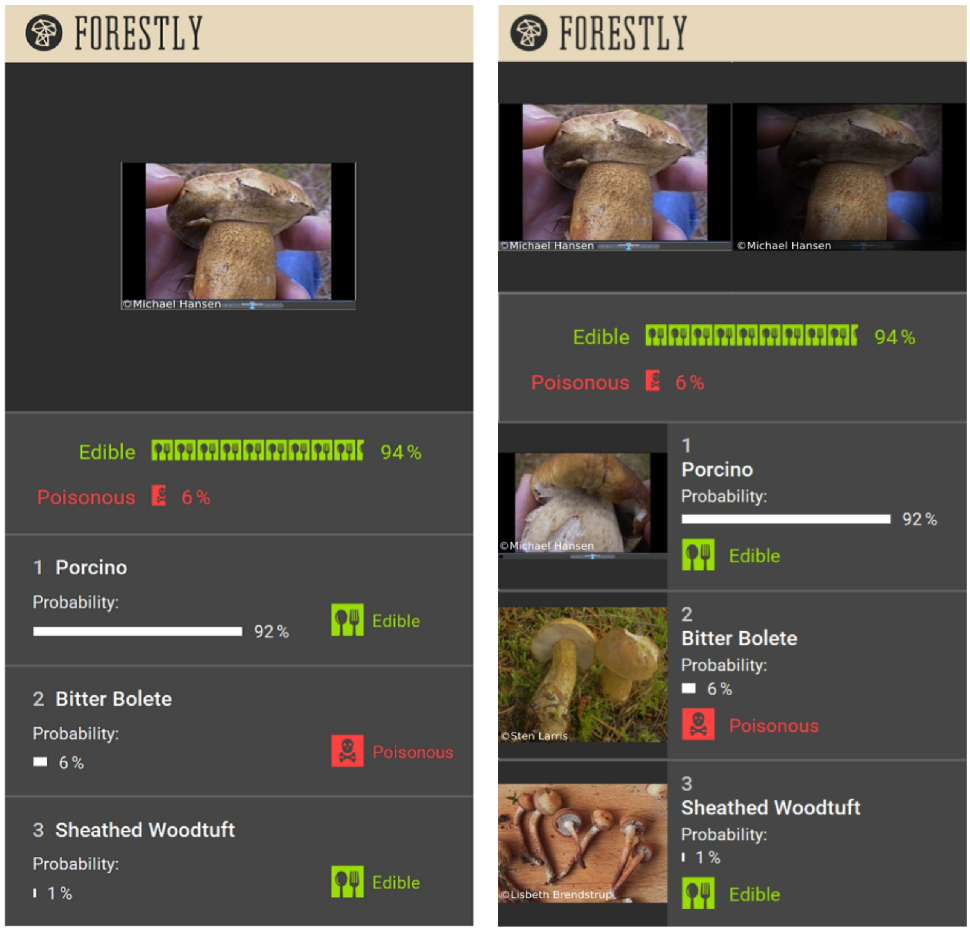
> "Assume you decide to take a mushroom with you based on a recommendation of the artificial intelligence and give it to a friend to eat. It turns out that it was a poisonous mushroom, and your friend complains of nausea, vomiting and diarrhea."

Participants then indicated how much blame they attributed to AI and to themselves for the friend being poisoned on two 7-point Likert scales ranging from (1) "no blame at all" to (7) "the greatest possible amount of blame" (see Komatsu et al., 2021; Malle et al., 2015).

Results and discussion

A Welch two-sample t-test showed that participants without visual explanations blamed AI significantly more ($M = 4.15$, $SD = 1.78$) than participants receiving visual explanations ($M = 3.72$, $SD = 1.68$) ($t(408) = 2.53$, $p = .012$) with a small effect size of $d = .25$ ($CI_{95} = [.05; .44]$). However, a second t-test showed that users do not significantly differ in self-blame between the groups without ($M = 5.34$, $SD = 1.50$) or with explanation ($M = 5.53$, $SD = 1.19$) ($t(393) = -1.45$, $p = .15$, $d = -.14$, $CI_{95} = [-.34; .05]$).

There could be several reasons for a significant difference in blaming AI. One reason is mind-attribution: The AI system is perceived as an actor with a mind of its own that (*i*) causally contributed to the negative consequence and (*ii*) would have been responsible and capable of preventing harm. A second reason could be "scapegoating", where the AI system is used as a target for blaming in order to reduce one's own guilt. In that case, it might be harder for people to externalize the blame to the AI for the group with



(a) No explanations (b) With explanations

Fig. 1 Examples of the two variants of the *Forestly* app interface for a correct prediction, (a) with AI decision only and (b) with additional visual explanations. Figure taken from Leichtmann, Humer, et al. (2023) licensed under CC-By Attribution 4.0 International

visual explanations, as their own capacity to detect the false classification is higher due to AI's irregularities explicitly shown in the explanations. Thus it is more difficult to use "scapegoating" as a strategy. The second result, indicating no difference between the groups in self-blame, however, does not suggest that either group attempted a stronger shift of blame away from oneself through a process of increased scapegoating.

To further investigate the underlying cognitive mechanisms of blame, we thus conducted a second study.

Study 2: exploring reasons for blame

Methods

We used data from $N = 501$ participants (240 female, 258 male, 2 non-binary, 1 without gender specification; mean age $M = 45.72$ years, $SD = 15.81$) collected in a second between-subjects online experiment, which explored the effects of various XAI methods on human decision-making and trust (Humer et al., 2024). The study design was similar to the previous study but (*i*) the decision tasks were embedded in a game, and (*ii*) we used a manipulation with four groups in which three XAI methods (E1, E2, E3, described in the following) were compared to a control group without explanations (C). Rather than a combination of GradCAM (E1) and example images (nearest neighbors, E2), the study tested the effects of both methods separately in two different groups. For a third XAI group (E3), we adapted the network dissection algorithm (Bau et al., 2020) to highlight the regions of each image where the most important detectors of the classifier were activated most strongly. We added textual labels for the concepts that the respective detector had learned to detect. The four different interface variants presented to the users are depicted in Fig. 2.

After 10 mushroom identification items, a vignette, conceptually similar to that in Study 1 (see Appendix A), was presented. Participants had to indicate how much blame they attributed to the AI-based system, the developers of AI, themselves, and the friend—each on a 7-point Likert scale as in the previous study. Next, participants indicated whether or not they thought that they could have recognized the wrong classification of the mushroom in this situation

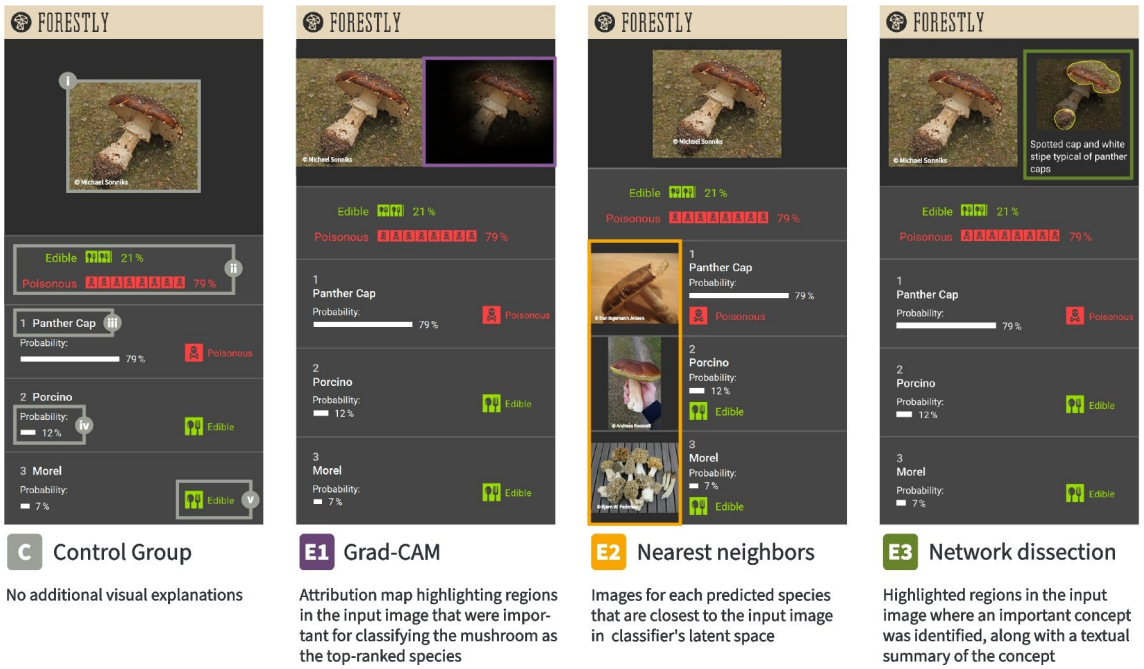


Fig. 2 Examples of the *Forestly* app interface for the four different groups of participants: An interface with the classification result but without explanations in the control group (C), and three interfaces with various explanations in the XAI groups using either Grad-CAM (E1), Nearest neighbors (E2), or network dissection (E3) as techniques. Figure taken from Humer et al. (2024) licensed under CC-By Attribution 4.0 International

(5-point Likert scale) and how severe they considered the consequences of poisoning (5-point Likert scale from "very mild" to "very severe"). Participants were then asked whether they thought that the AI-based system (*i*) had free will, (*ii*) was responsible for its actions, (*iii*) acted with intention, (*iv*) had a mind of its own, and (*v*) had the ability to make decisions on its own (5-point Likert scales from "strongly disagree" to "strongly agree")—common constructs in moral reasoning (e.g., Bigman et al., 2019; Shank and DeSanti, 2018; Komatsu et al., 2021; Monroe et al., 2017; Malle et al., 2015).

Results and discussion

We tested the effect of visual explanations on the attribution of blame by comparing blame ratings from each of the test groups (i.e., the groups receiving one of the explanation techniques E1, E2, and E3) to the control group (i.e., the group receiving the classification outcome only, C) in many-to-one comparisons. Results of the Dunnett tests indicate that participants with explanations (E1, E2, E3) did not blame AI (see Table 1) or themselves (see Table 2) statistically differently compared to the control group (C), although there is a tendency to blame oneself more when receiving

Table 1 Many-to-one comparisons of groups with explanations to the control group on *blaming the AI*

	Control Group		Test Group		Dunnett			Welch			Effect Size		
Meth.	*M*	*SD*	*M*	*SD*	*SE*	*t*	*p*	*t*	*df*	*p*	*d*	*CI* Lo	*CI* Hi
E1	4.72	1.53	4.53	1.72	0.195	-0.99	0.65	0.96	233	0.17	-0.12	-0.36	0.12
E2	4.72	1.53	4.62	1.60	0.198	-0.49	0.93	0.50	233	0.31	-0.06	-0.31	0.18
E3	4.72	1.53	4.60	1.53	0.195	-0.64	0.87	0.66	250	0.25	-0.08	-0.32	0.16

Table 2 Many-to-one comparisons of groups with explanations to the control group on *blaming oneself*

	Control Group		Test Group		Dunnett			Welch			Effect Size		
Meth.	*M*	*SD*	*M*	*SD*	*SE*	*t*	*p*	*t*	*df*	*p*	*d*	*CI* Lo	*CI* Hi
E1	4.61	1.32	4.71	1.64	0.175	0.59	0.90	-0.56	218	0.29	0.07	-0.17	0.31
E2	4.61	1.32	4.95	1.41	0.178	1.91	0.15	-2.00	229	0.02	0.25	0.01	0.50
E3	4.61	1.32	4.74	1.38	0.175	0.73	0.81	-0.78	244	0.22	0.10	-0.15	0.34

the example-based explanation (E2). We thus could not replicate the findings of the previous study.

While explanations did not cause statistically significant differences in blaming, we found differences depending on the target of blame (i.e., the actors involved in the scenario, directly and indirectly; see *M* and *SD* in Table 3). Participants attributed most blame to themselves and the AI system with no statistically significant difference ($t_{\text{AI–self}}(989.5) = 1.15,\ p = 0.25$). These two ratings significantly differed from the lower blame ratings of the developer ($t_{\text{self–developer}}(977.6) = 4.77,\ p < .001$; $t_{\text{AI–developer}}(997.6) = 3.49,\ p < .001$). The lowest amount of blame was attributed to the friend ($t_{\text{developer–friend}}(992.3) = 5.58,\ p < .001$).

To better understand the causes for blame we calculated correlation coefficients of blame attributions and various variables typically involved in moral cognition (see Table 3). The results reveal that blaming, in general, is associated with the severity of the consequences (see 7. in Table 3). Additionally, blaming the AI system is correlated with ascribing responsibility to the AI system. This responsibility is in turn associated with various aspects of mind perception (i.e., the ability to make its own decisions, having mind, having free will; all inter-correlated within a range of.33 to.64)—making the AI system an actor on its own. For a better overview, these relationships are graphically illustrated as a network model in Fig. 3.

Interestingly, blaming AI and blaming oneself is uncorrelated in this study. However, a medium correlation was found for blaming AI and blaming the developers ($r = .63,\ p < .001$).

To sum up the quantitative results of study 2, we did not find any group differences in blaming AI or oneself. This is somehow contradicting our previous results and also the theoretical idea that AI-based systems explaining their classification results would be blamed differently either because (*i*) it is perceived as having a mind and thus capable of blame, or (*ii*) higher traceability through explanations does not allow to externalize ones' own blameworthiness (scapegoating).

However, the differences in the allocation of blame between the actors allow for a conclusion to be drawn. The actors with the closest distance to the main-cause event received the most blame, and more distanced actors with indirect influence such as the developer received lower blame. In line with the moral typecasting theory Gray and Wegner 2009, the victim of the event received the lowest amount of blame. This confirms the Path Model of Blame according to Malle et al. (2014) insofar as those who had a direct influence on the initial cause of the negative outcome are blamed more.

Further evidence to confirm the path model can also be found in the correlative findings. First, severity is associated with blaming behavior. This symbolizes the first decision of the blame model by Malle et al. (2014), which requires an event of a certain severity. Furthermore, correlations confirm the role of perceived agency in moral perceptions in the context of XAI. According to the Path Model of Blame, ascribed obligation and capacity to prevent an event are important components in the cognitive process of blaming. This role of capacity is also evident in the degree to which participants blame themselves for the negative event, as it is predicted by the degree to which they think they could have recognized the wrong classification.

Surprisingly, hints for scapegoating are limited in this study. Blaming AI and blaming oneself is uncorrelated contradicting the idea that AI was used as a scapegoat in order to lower self-blame. Although participants also seemed to

Table 3 Descriptive statistics and correlations for study variables

	M	*SD*	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
1. Blaming AI[a)]	4.63	1.59	—										
2. Blaming Self[a)]	4.74	1.43	0.01	—									
3. Blaming Developers[a)]	4.27	1.67	0.63***	0.01	—								
4. Blaming Friend[a)]	3.70	1.52	-0.15*	-0.01	-0.14	—							
5. Opportunity to Prevent[b)]	2.67	0.98	-0.05	0.17**	0.00	0.06	—						
6. AI Responsibility[b)]	2.36	1.15	0.30***	-0.11	0.37***	-0.14	0.10	—					
7. Poisoning Severity[b)]	3.45	0.74	0.16*	0.25***	0.16*	-0.06	0.02	0.05	—				
8. AI Free Will[b)]	1.80	1.00	0.02	-0.02	0.12	0.00	0.18	0.38***	0.03	—			
9. AI Intentionality[b)]	1.79	1.01	0.09	0.00	0.19***	0.09	0.12	0.43***	0.01	0.49***	—		
10. AI Own Mind[b)]	1.99	1.09	0.01	-0.04	0.07	-0.07	0.14	0.37***	0.06	0.64***	0.49***	—	
11. AI Own Decisions[b)]	2.69	1.21	0.04	-0.09	0.06	-0.02	0.02	0.33***	-0.05	0.36***	0.33***	0.44***	—

[a)]range 1 to 7 Likert-scale level

[b)]range 1 to 5 Likert-scale level

$*\ p_{\text{adj}} < .05;\ **\ p_{\text{adj}} < .01;\ ***\ p_{\text{adj}} < .001$

Significance levels corrected for multiple testing

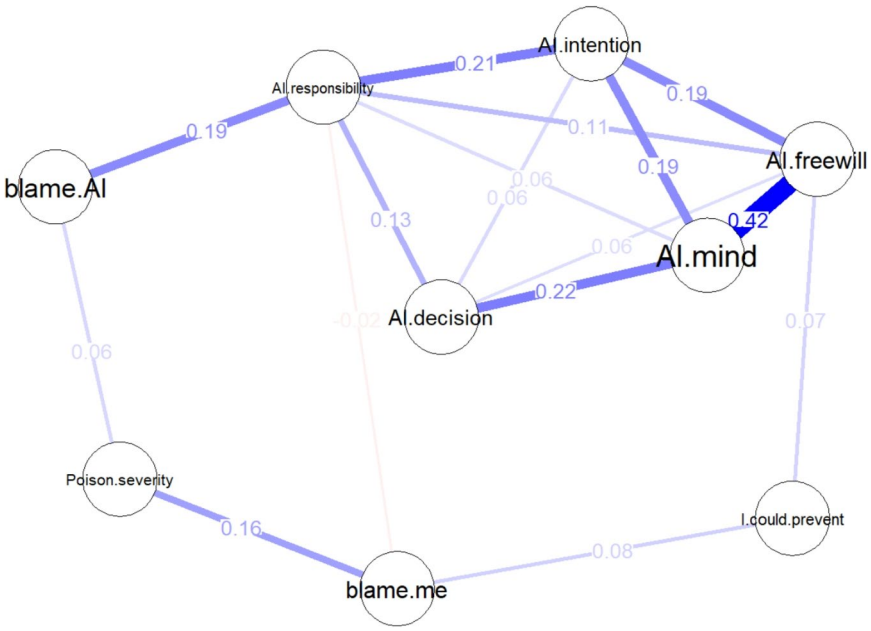


Fig. 3 Network model of moral reasoning in the fictitious scenario. Variables (i.e., moral-cognitive responses to the moral event and the actors) are represented as nodes and relationships between the variables are depicted as edges. Larger weights are doubly encoded by shorter edge distance and thicker line width. Note: Blaming variables are the degree to which participants blamed AI (*blame.AI*) and themselves (*blame.me*); *poison.severity* is the perceived severity of the harm; AI-related variables are the perceived responsibility of the AI system (*AI.responsibility*), and the degree to which the AI system is perceived as having intentions (*AI.intention*),a free will (*AI.freewill*), a mind on its own (*AI.mind*), and the ability to make its own decisions (*AI.decision*); *I.could.prevent* is the degree to which participants have the impression they could have recognized the wrong classification

blame the developers as the humans behind the technology (indicated by the medium correlation between blaming AI and blaming the developers), they did not seem to serve as scapegoats either.

For a more profound insight into the cognitive processes of blame attribution, qualitative content analysis was used to examine responses provided in freeform text fields. It can be concluded that most people did not assume intention in poisoning the friend. Thus, most reasons for attributing blame are based on opinions of the obligation or capacity to *prevent* the food poisoning. For example, a person blaming themselves argued "[...] I could also have had it [i.e., the mushroom] evaluated by an expert [...]", indicating that they have the obligation and capacity to prevent the event (i.e., by double-checking the classification). Similar reasons were mentioned for blaming AI ("because apparently it must consider even more aspects of these mushrooms"), the developers ("developers must check their app [...]"), or the friend ("because he himself should have checked what he eats"). However, some people stopped in their reasoning on a previous step: They identified all actors in the causal chain of events and then attributed a similar amount of blame to every actor without differentiation ("Everyone who contributes to something happening is partly to blame"). Thus, for some, everyone or no one was blameworthy (e.g., if people believed in coincidences: "No one is to blame. It is an accident."). While most participants blamed most actors to at least some degree, only few individuals seemed to have explicitly used strategies of scapegoating, such as portraying themselves as a "victim" of another actor's action ("[the AI] told me to pick the mushroom").

General discussion and future work

In two empirical studies we explored the effects of system transparency (i.e., XAI methods) on human blaming behavior, the distribution of blame, and underlying moral reasoning in scenarios where a human is harmed as a result of human–AI interaction using well-established mushroom picking tasks (e.g., Leichtmann, Hinterreiter, et al., 2023; Leichtmann, Humer, et al., 2023; Humer et al., 2024).

While the first study showed that humans who did not receive explanations blamed the AI-based system more than participants who received a combination of two XAI methods, we could not reproduce these findings in a second study that investigated three distinct XAI methods. It could be argued that explanation methods vary in their effects depending on various characteristics of these methods. For example, some explanations might hint toward a wrong classification due to mismatching information. This could mean that some explanation methods lead to higher traceability than other methods. Although individuals may have tended to blame a system lacking explanations (i.e., using it as scapegoat), this strategy may not be as applicable with systems offering greater traceability, as users could have better identified a fault within the system. However, not every XAI method may equally contribute to enhanced traceability, leading to varying effects. Future studies will have to test the effect of example-based XAI methods with larger sample sizes to confirm this. Additionally, users' perceptions of traceability within AI systems (see Schrills and Franke, 2023) and whether this perception influences their tendency to attribute blame, could provide valuable insights in future studies.

In our study, we used explanations that Miller (2023) calls 'recommend and defend' approaches. The AI decision aid gives a recommendation and explains why this recommendation is considered the best answer. In such high levels of automation of decision making the user is usually not able to explore different options (Parasuraman et al., 2000). Therefore, these explanations offer little help when the user does not find the recommendation convincing (Miller, 2023). They do not provide information on why other solutions are not considered the best answers, as counterfactual explanations do. Future studies could compare differences in blaming between these two approaches. Users might be more likely to blame the system if they feel it is defending its decisions as in 'recommend and defend' approaches instead of being transparent and understandable. On the other side, the possibility to ask why other recommendations were not chosen in counterfactual explanations could *(i)* make it easier to recognize when the system makes mistakes, *(ii)* increase the locus of control of the users, and thus *(iii)* make the users more responsible. This way, the AI decision aid cannot be easily used as a scapegoat. Users have to take more responsibility themselves and might be more blameworthy.

Besides XAI methods, there are other features of the interface design that might affect user blaming behavior. Our study focused on XAI methods, mostly in the form of visual explanations, that is to provide information about the causal history of the AI classification formation in visual form (for a definition of explanation see Miller, 2019). Future studies should test the influence of other design features on blaming behavior beyond explanations. One such aspect that has been studied in the AI literature could be the system's confidence level (see Zhang et al., 2020). In our study, we ensured that this level varied across task items (indicated here in different percentages between 40% to 100% for the top classes) to evoke an overall balanced level of certain and uncertain decisions. Future studies could specifically manipulate this certainty information and investigate the effect on blaming behavior. Users might be even more inclined to blame an AI-based app with higher confidence levels and even feel deceived by the AI-based system.

Our study shows that some design features of AI-based systems, such as XAI methods, can affect our moral reasoning. Therefore, these effects must be considered in the development of AI-based systems. Future studies will have to test the effects of different variations and combinations of interface features including other XAI methods or variations in certainty levels.

The second study allowed us to draw conclusions about moral reasoning: Correlations showed that people blamed AI more if it was perceived as responsible for its actions. This responsibility in turn was associated with mind perception, in line with work by Gray et al. (2012). Furthermore, people blamed themselves more if they thought they could have recognized the false classification. This is also supported by a qualitative analysis of participants' comments as they justify their blaming mainly with an actor's obligation or capacity to prevent an event.

These findings reinforce, on the one hand, the significance of both the capability and obligation of every agent in blame attribution and confirm the Path Model of Blame (Malle et al., 2014). They also affirm the idea that mind perception plays a certain role in attributing blame to machines, albeit only indirectly. Future studies, however, need to examine more closely the extent to which XAI methods influence such mind perception, and if they do influence it, the extent to which these effects counteract other effects due to increased traceability. At this point, it is important to note that the results concerning the cognitive processes of blaming behavior are only of a correlational nature. They are causally described here only in the light of the Path Model of Blame that indicates a certain causal chain (Malle et al., 2014). However, it should be emphasized that the data does not allow to draw conclusions on causal relationships, but represent correlative associations.

We found only limited evidence for "scapegoating". In the first study, it was evident that AI-based systems without explanations were attributed more blame. This can be explained through scapegoating, as with a system providing explanations, there is a greater chance of recognizing a flaw in the decision, making the AI-based system less likely to be used as a scapegoat for blame. However, it could not be demonstrated that this also led to a significantly higher level of self-blame. Qualitative results also yielded limited evidence of scapegoating processes. Future studies should thus focus more on the circumstances under which scapegoating is used as a strategy.

Finally, our study also showed that blame is not distributed equally. People blamed themselves and the AI-based system most, followed by the developers. The friend received the lowest amount of blame. Future studies are planned to explore whether this blame of each actor is also associated with various behavioral intentions. For example, blaming oneself might be associated with the intention of informing oneself better about the limitations of AI, while blaming developers might be linked to potential legal actions against a company.

Appendix A Vignette of study 2

The vignette presented in Study 2 was similar to that presented in Study 1, but with small variations in wording:

> "Imagine you had been on a real-life mushroom hunt using the "Forestly" app from this study. Assume that

you decided to pick a mushroom which had been classified as edible by the artificial intelligence. After your mushroom hunt, you met a friend of whom you know that he is very into mushrooms. You decide to give him this mushroom as a gift. In the evening it turns out that the mushroom was poisonous and your friend complains about nausea, vomiting, and diarrhea."

Funding Open access funding provided by Johannes Kepler University Linz. This work was funded by Johannes Kepler University Linz, Linz Institute of Technology (LIT), the State of Upper Austria, and the Federal Ministry of Education, Science and Research under grant number LIT-2019-7-SEE-117, awarded to MM and MS, the Austrian Science Fund under grant number FWF DFH 23-N, and under the Human-Interpretable Machine Learning project (funded by the State of Upper Austria).

Data availibility The datasets generated and analyzed during the current study are available on OSF at https://osf.io/375xu/.

Declarations

Conflicts of interest The authors have no competing interests to declare.

Ethics approval Both studies complied with the tenets of the Declaration of Helsinki and adhered to ethical guidelines of the APA Code of Conduct, data protection regulations in Europe, as well as local legislation and institutional requirements. Informed consent was obtained from each participant prior to data collection. Participation was voluntary and could have been terminated at any point without consequences.

References

Barredo Arrieta, A., & Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bastian, B., Laham, S. M., Wilson, S., Haslam, N., & Koval, P. (2011). Blaming, praising, and protecting our humanity: The implications of everyday dehumanization for judgments of moral status: Dehumanization and moral status. *British Journal of Social Psychology, 50*(3), 469–483. https://doi.org/10.1348/014466610X521383

Bau, D., Zhu, J.-Y., Strobelt, H., Lapedriza, A., Zhou, B., & Torralba, A. (2020). Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences, 117*(48), 30071–3007. https://doi.org/10.1073/pnas.1907375117

Bigman, Y. E., Waytz, A., Alterovitz, R., & Gray, K. (2019). Holding robots responsible: The elements of machine morality. *Trends in Cognitive Sciences, 23*(5), 365–368. https://doi.org/10.1016/j.tics.2019.02.008

Brandenburg, W. E., & Ward, K. J. (2018). Mushroom poisoning epidemiology in the United States. *Mycologia, 110*(4), 637–641. https://doi.org/10.1080/00275514.2018.1479561

Cervellin, G., Comelli, I., Rastelli, G., Sanchis-Gomar, F., Negri, F., De Luca, C., & Lippi, G. (2018). Epidemiology and clinics of mushroom poisoning in Northern Italy: A 21-year retrospective analysis. *Human & Experimental Toxicology, 37*(7), 697–703. https://doi.org/10.1177/0960327117730882

Copp, C. J., Cabell, J. J., & Kemmelmeier, M. (2023). Plenty of blame to go around: Attributions of responsibility in a fatal autonomous vehicle accident. *Current Psychology, 42*(8), 6752–6767. https://doi.org/10.1007/s12144-021-01956-5

Gerstenberg, T., & Lagnado, D. A. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition, 115*(1), 166–17. https://doi.org/10.1016/j.cognition.2009.12.011

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619–619. https://doi.org/10.1126/science.1134475

Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology, 96*(3), 505–520. https://doi.org/10.1037/a0013748

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

Greene, J. D. (2015). The rise of moral cognition. *Cognition, 135*, 39–42. https://doi.org/10.1016/j.cognition.2014.11.018

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 1–4. https://doi.org/10.1145/3236009

Hong, J.-W., Wang, Y., & Lanz, P. (2020). Why is artificial intelligence blamed more? Analysis of faulting artificial intelligence for self-driving car accidents in experimental settings. *International Journal of Human-Computer Interaction, 36*(18), 1768–1774. https://doi.org/10.1080/10447318.2020.1785693

Humer, C., Hinterreiter, A., Leichtmann, B., Mara, M., & Streit, M. (2024). Reassuring, misleading, debunking: Comparing effects of XAI methods on human decisions. *ACM Transactions on Interactive Intelligent Systems, 3665647*. https://doi.org/10.1145/3665647

Jeyakumar, J. V., Noor, J., Cheng, Y.-H., Garcia, L., & Srivastava, M. (2020). How can i explain this to you? An empirical study of deep neural network explanation methods. *Advances in Neural Information Processing Systems, 33*.

Ketelaar, T., Tung, & Au., W. (2003). The effects of feelings of guilt on the behaviour of uncooperative individuals in repeated social bargaining games: An affect-as-information interpretation of the role of emotion in social interaction. *Cognition and Emotion, 17*(3), 429–453. https://doi.org/10.1080/02699930143000662

Kim, T., & Hinds, P. (2006). Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 80–85). Univ. of Hertfordshire, Hatfield, UK: IEEE. https://doi.org/10.1109/ROMAN.2006.314398

Komatsu, T., Malle, B.F., & Scheutz, M. (2021). Blaming the reluctant robot: Parallel blame judgments for robots in moral dilemmas across U.S. and Japan. *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 63–72). Boulder CO USA: ACM. https://doi.org/10.1145/3434073.3444672

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., & Baum, K. (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*, 103473. https://doi.org/10.1016/j.artint.2021.103473

Leichtmann, B., Hinterreiter, A., Humer, C., Streit, M., & Mara, M. (2023). Explainable Artificial Intelligence improves human decision-making: Results from a mushroom picking experiment at a public art festival. *International Journal of Human-Computer Interaction, 1–18*. https://doi.org/10.1080/10447318.2023.2221605

Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. (2023). Effects of explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior, 139*, 10753. https://doi.org/10.1016/j.chb.2022.107539

Malle, B.F. (2019). How many dimensions of mind perception really are there? A.K. Goel, C.M. Seifert, and C. Freska (Eds.), *Proceedings of the 41st annual meeting of the cognitive science society* (pp. 2268–2274). Montreal, Canada: Cognitive Science Society.

Malle, B.F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 117–124). Portland Oregon USA: ACM.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*(2), 147–186. https://doi.org/10.1080/1047840X.2014.877340

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Miller, T. (2023). Explainable AI is dead, long live explainable AI!: Hypothesis-driven decision support using evaluative AI. *2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 333–342). Chicago IL USA: ACM. https://doi.org/10.1145/3593013.3594001

Molnar, C. (2023). *Interpretable machine learning*. Retrieved 2022-09-12, from https://christophm.github.io/interpretable-ml-book/

Monroe, A. E., Brady, G. L., & Malle, B. F. (2017). This isn't the free will worth looking for: General free will beliefs do not influence moral judgments, agent-specific choice ascriptions do. *Social Psychological and Personality Science, 8*(2), 191–199. https://doi.org/10.1177/1948550616667616

Nelissen, R., Dijker, A., & deVries, N. (2007). How to turn a hawk into a dove and vice versa: Interactions between emotions and goals in a give-some dilemma game. *Journal of Experimental Social Psychology, 43*(2), 280–28. https://doi.org/10.1016/j.jesp.2006.01.009

Parasuraman, R., Sheridan, T., & Wickens, C. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 30*(3), 286–29. https://doi.org/10.1109/3468.844354

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

Rothschild, Z. K., Landau, M. J., Sullivan, D., & Keefer, L. A. (2012). A dual-motive model of scapegoating: Displacing blame to reduce guilt or increase control. *Journal of Personality and Social Psychology, 102*(6), 1148–116. https://doi.org/10.1037/a0027413

Schmutz, M., Carron, P.-N., Yersin, B., & Trueb, L. (2018). Mushroom poisoning: a retrospective study concerning 11-years of admissions in a Swiss Emergency Department. *Internal and Emergency Medicine, 13*(1), 59–67. https://doi.org/10.1007/s11739-016-1585-5

Schrills, T., & Franke, T. (2023). How do users experience traceability of AI systems? Examining subjective information processing awareness in Automated Insulin Delivery (AID) systems. *ACM Transactions on Interactive Intelligent Systems, 13*(4), 1–34. https://doi.org/10.1145/3588594

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626). Venice: IEEE. https://doi.org/10.1109/ICCV.2017.74

Shank, D. B., & DeSanti, A. (2018). Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior, 86*, 401–411. https://doi.org/10.1016/j.chb.2018.05.014

Sullivan, Y. W., & Fosso Wamba, S. (2022). Moral judgments in the age of Artificial Intelligence. *Journal of Business Ethics, 178*(4), 917–943. https://doi.org/10.1007/s10551-022-05053-w

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences, 14*(8), 383–388. https://doi.org/10.1016/j.tics.2010.05.006

Wischnewski, M., Krämer, N., Müller, E. (2023). Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Hamburg Germany: ACM. https://doi.org/10.1145/3544548.3581197

Yang, F., Huang, Z., Scholtz, J., Arendt, D.L. (2020). How do visual explanations foster end users' appropriate trust in machine learning? *Proceedings of the 25th International Conference on Intelligent User Interfaces* (pp. 189–201). Cagliari Italy: ACM. https://doi.org/10.1145/3377325.3377480

Zhang, Y., Liao, Q.V., Bellamy, R.K.E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 295–305). Barcelona Spain: ACM. Retrieved 2024-02-23, from https://dl.acm.org/doi/10.1145/3351095.3372852