Towards a computational model of responsibility judgments in sequential human-AI collaboration

Stratis Tsirtsis[1], **Manuel Gomez-Rodriguez**[1], **Tobias Gerstenberg**[2]
[1]{stsirtsis,manuelgr@mpi-sws.org} Max Planck Institute for Software Systems, Germany
[2]{gerstenberg@stanford.edu} Department of Psychology, Stanford University, USA

Abstract

When a human and an AI agent collaborate to complete a task and something goes wrong, who is responsible? Prior work has developed theories to describe how people assign responsibility to individuals in teams. However, there has been little work studying the cognitive processes that underlie responsibility judgments in human-AI collaborations, especially for tasks comprising a sequence of interdependent actions. In this work, we take a step towards filling this gap. Using semi-autonomous driving as a paradigm, we develop an environment that simulates stylized cases of human-AI collaboration using a generative model of agent behavior. We propose a model of responsibility that considers how unexpected an agent's action was, and what would have happened had they acted differently. We test the model's predictions empirically and find that in addition to action expectations and counterfactual considerations, participants' responsibility judgments are also affected by how much each agent actually contributed to the outcome.

Keywords: responsibility, counterfactual simulation, sequential decision making, human-AI collaboration

Introduction

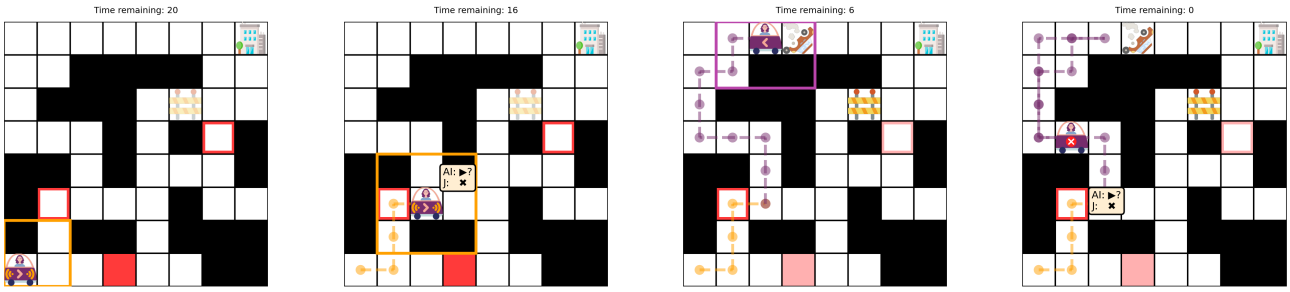
Imagine a future where every car is supported by an AI agent with autonomous driving capabilities. Jane starts driving her car manually, she makes a quick stop to leave her kids at school, and then she enters her car again. Alan, her AI driving assistant, offers to drive her to work, she accepts, and she relaxes while enjoying the ride. Alan follows a path, different from the one she would have followed, that seems to have less traffic than usual. Unbeknownst to both of them, there is a car crash blocking a street and they need to turn around and find another (longer) way. As a result, Jane arrives late at work. Who is responsible for the delay? Alan for taking a path that was blocked or Jane for letting the AI drive in the first place? Both of them, since they both drove part of the commute, or none of them, since they didn't know about the accident?

Questions about responsibility are ubiquitous in our everyday lives and humans make responsibility judgments intuitively even about complex situations such as the one described above. Cognitive scientists have developed and tested different theories about the cognitive process underpinning responsibility judgments (Alicke, 2000; Chockler & Halpern, 2004; Gerstenberg & Lagnado, 2010; Shaver, 2012). However, the increasing development of AI systems that *assist* and *collaborate* with humans, rather than replacing them (Balazadeh Meresht et al., 2022; De et al., 2020, 2021; Mozannar et al., 2022; Okati et al., 2021; Raghu et al., 2019; Straitouri et al., 2021; Wilder et al., 2021), calls for more empirical and theoretical research to shed light on the way humans make responsibility judgments in situations involving human-AI teams (Cañas, 2022). Recent work in that area has identified several factors that influence responsibility judgments (Awad et al., 2020; Lima et al., 2021; Longin et al., 2023). However, this work has not attempted to characterize the underlying cognitive processes that support such judgments. Our work takes a step towards filling this gap by introducing a computational model to predict responsibility judgments for human-AI teams in environments where the two agents collaborate and act sequentially.

Responsibility & counterfactual reasoning

Existing theories about the cognitive process of responsibility attribution have established strong ties with causality (Pearl, 2009) and counterfactual reasoning (Byrne, 2016; Kahneman et al., 1982; Roese, 1997). Humans tend to consider an object, event, action or agent as (causally) responsible for an outcome if they can mentally simulate an alternative reality where that outcome would have been different if the candidate cause had not existed or occurred in the first place (Beckers, 2023; Chockler & Halpern, 2004; Gerstenberg et al., 2018; Halpern & Kleiman-Weiner, 2018; Lagnado et al., 2013; Langenhoff et al., 2021; Triantafyllou et al., 2022; Wu & Gerstenberg, 2024; Wu et al., 2023; Xiang et al., 2023; Zultan et al., 2012). In that context, Gerstenberg et al. (2021) have developed the counterfactual simulation model (CSM), a computational model that accurately predicts the extent to which people perceive an object (*e.g.*, a moving billiard ball) as a cause of an observed outcome (*e.g.*, potting another ball). Specifically, using a physics engine to approximate people's intuitive understanding of physics (Gerstenberg & Tenenbaum, 2017; Ullman et al., 2017), the model performs (stochastic) simulations of counterfactual situations where the candidate cause (*e.g.*, the moving billiard ball) is removed from the scene or slightly perturbed. Then, it predicts participants' causal judgments based on the estimated probability that the outcome would have been different had the respective intervention on the candidate cause taken place.

More recently, Wu et al. (2022, 2023) have explored extensions of the CSM in social settings using Markov decision processes (MDPs) (Sutton & Barto, 2018) as generative models of agent behavior. Reminiscent of the results in the phys-



(a) The **AI** starts driving, unaware of the road closure

(b) The **AI** asks for confirmation to go right and **Jane** rejects

(c) **Jane** takes control of the car but encounters an accident

(d) Time runs out and they **fail** to reach the workplace

Figure 1: **Illustration of a commute in our semi-autonomous driving environment.** The human agent (**Jane**) and the **AI** are both in the same car and their goal is to reach the workplace within the time limit shown above the grid. The sign indicates that the AI is in control. The grid contains three traffic spots, one congested (■) and two non congested (□), whose status is initially known only to the AI. It also contains a road closure () which is known to the human but unknown to the AI. Obstacles that are unknown to the agent in control but known to the other agent appear faded. The arrow signs marked on the car (*e.g.*,) indicate the direction that the driver in control is planning to follow. The 3×3 rectangle around the car represents the agents' field of view via which they discover obstacles that are previously unknown to them. Here, the accident () present at the top row of the grid becomes visible only after the car goes next to it and it enters the agent's field of view.

ical domain, they have shown that the CSM predicts people's judgments about the extent that a decision of a psychological agent caused an outcome based on counterfactual simulations where that agent has made a different decision (Wu et al., 2022). However, in the context of responsibility attribution, the shift of focus from physical objects to agents introduces additional complexity, since an agent's actions are conditioned on their epistemic state (Beckers, 2023; Franklin et al., 2022; Halpern & Kleiman-Weiner, 2018; Kirfel & Lagnado, 2021). To explore this further, Wu et al. (2023) have experimented with a gridworld environment where an agent is trying to achieve an outcome in the presence of a second (potentially adversarial) agent. They have proposed an extension of the CSM that additionally models the first agent's *belief* about the second agent's intention and explains responsibility judgments by combining counterfactual simulations with intention inferences (Kleiman-Weiner et al., 2015).

Our contributions

We further extend the CSM by developing and experimenting with a stylized but rich semi-autonomous driving environment, where a human and an AI agent collaborate towards a common objective. A distinctive feature of the setting we focus on is that the two agents share the same goal but have partial and differing knowledge about elements of the physical environment they operate in. As a result, they hold different beliefs about the state of the world, which they update either via direct observations or via inferences from each other's actions (Baker et al., 2009). Moreover, the two agents take a series of interdependent actions, and their relationship is asymmetric, with the human having (some) control over the actions of the AI which, in turn, plays an assistive role. We propose a model of responsibility for the human and the AI that relies on counterfactual simulations to estimate how unexpected an agent's action was, and what would have happened had each agent acted differently. In an online experiment, we find that participants' responsibility judgments about the human are affected by counterfactuals and are well-captured by our model. On the other hand, a simpler model based solely on the actual contribution to the outcome captures responsibility judgments about the AI.

Computational model

We develop a 2D gridworld environment that simulates and illustrates stylized cases of commute.[1] Below, we start by providing a high-level description of our environment. Then, we formalize its main elements, and we introduce a generative model of agent behavior. Building upon that, we propose a model to predict responsibility judgments about the human and the AI agent in individual commutes.

Environment description

Consider the illustration in Figure 1: The two agents (human & AI) are in a car, which is initially placed at the bottom left corner of an 8×8 grid consisted of black and white (road) tiles. The grid is known to both agents a priori and they both share a common goal – to reach the human's workplace at the top right corner within a given time limit. The simulation proceeds in time steps and, at each time step, the car is controlled either by the AI or the human. The agent who is in control can move the car horizontally or vertically by one tile per time step. Moving to a tile is possible only if it is white (*i.e.*, a road) and it is not blocked by a road closure or an accident. The grid may also contain traffic spots that are either

[1] Our code and data are accessible at https://github.com/cicl-stanford/responsibility_sequential

congested or not congested for the entire commute, with congested ones causing the car to remain idle for 10 time steps.

Each agent has only partial knowledge of potential obstacles in the environment. The human knows about road closures and the locations of the traffic spots but not about their congestion status. The AI knows everything about traffic spots but it is unaware of road closures. Lastly, accidents may appear randomly on any tile, and they are unknown to both of them. Each agent discovers a previously unknown obstacle only once it enters their field of view surrounding the car.

The two agents collaborate with each other by switching control of the car. One of them starts driving and, at a randomly chosen time step, the AI asks the human whether they want to switch control for the remainder of the commute. If the AI is driving, it requests confirmation to continue; if the human is driving, the AI asks whether it should take control of the car. The human decides based on the information they have about the environment at the time, and we will refer to this decision as the *switching decision*. The agent who is in control after that point drives until they reach the workplace (success) or until time runs out (failure).

Formal framework

Our environment can be described using the framework of decentralized partially observable MDPs (Bernstein et al., 2002; Oliehoek, Amato, et al., 2016; Triantafyllou et al., 2022). Therein, an episode unfolds over T time steps (here, the time limit to reach the workplace) and includes more than one agent (here, the human and the AI) who act independently. At each time step t, the process is characterized by a state $\boldsymbol{s}_t \in \mathcal{S}$ and, in our case, contains information about the world such as the location of the car and the identity of the current driver. The two agents take actions $a_{H,t} \in \mathcal{A}_H$, $a_{AI,t} \in \mathcal{A}_{AI}$, that correspond to doing nothing, moving on the grid, offering or accepting/rejecting to switch control and combinations thereof. A function $f_S : \mathcal{S} \times \mathcal{A}_H \times \mathcal{A}_{AI} \to \mathcal{S}$ controls the (deterministic) transitions between states and, at each time step, the agents receive a numerical reward – a positive value if the car has reached the workplace and -1 otherwise. Their goal is to maximize their total reward. Moreover, each agent is characterized by a belief P_{agent} about the state of the world and takes actions a sampled from a (stochastic) policy $\pi_{agent}(a \,|\, P_{agent})$.

Beliefs & observations Here, we focus on the agents' beliefs and their (partial) observability model, which form the basis for our generative model of agent behavior and the responsibility model we present next. The two agents start with their own prior beliefs, formalized as two distributions P_H, P_{AI} over all states in $\mathcal{S}$, where the uncertainty originates from their partial knowledge about obstacles (*i.e.*, traffic spots, road closures, accidents) that may be present on the grid.

Since the human is aware of road closures, their prior belief has zero probability on states $\boldsymbol{s}$ whose road closures do not match with the true state $\boldsymbol{s}_0$. Moreover, since accidents are unexpected, we set the prior probability of any state that contains an accident to a negligible amount close to zero. To model the human's ignorance about the congestion status of K usual traffic spots in the grid, we set their prior uniformly over states corresponding to the 2^K different combinations of congestion status. The AI's prior is defined in a similar way, ensuring that the AI knows the true congestion status of traffic spots but ignores potential road closures and accidents.

At each time step, the two agents receive an observation $\boldsymbol{o}_t = \text{FOV}(\boldsymbol{s}_t)$ that includes all the obstacles within their field of view. Based on this observation, both agents update their beliefs about the state of the world by eliminating any state that would contradict their field of view, that is,

$$P_{agent}(\boldsymbol{s} \,|\, \boldsymbol{o}_t) \propto \mathbb{1}\left[\boldsymbol{o}_t = \text{FOV}(\boldsymbol{s})\right] \cdot P_{agent}(\boldsymbol{s}) \;\; \forall \boldsymbol{s} \in \mathcal{S},$$

where $\mathbb{1}[\cdot]$ denotes the indicator function. Moreover, whenever the AI is in control of the car, the human receives an enhanced observation $\boldsymbol{o}_t = (\text{FOV}(\boldsymbol{s}_t), a_{AI,t})$ that also includes the AI's action. Motivated by prior work that models action understanding as Bayesian inverse planning (Baker et al., 2009, 2017), we assume that they update their belief about the congestion status of the traffic spots based on the direction that the AI intends to move. Let $a_{AI,t} = d$ denote a movement in direction d (*e.g.*, $d = \text{LEFT}$) and π_H be the human's policy. The human performs a Bayesian update on their belief by considering the likelihood that they would have chosen direction d if they had the same belief as the AI. Formally, let $\tilde{P}_{AI}$ be a function that takes as input a state $\boldsymbol{s}$ and returns a belief (*i.e.*, a distribution over states) oblivious to any road closures in $\boldsymbol{s}$ that have not yet entered the agents' field of view. The human's Bayesian update, as described above, takes the form

$$P_H(\boldsymbol{s} \,|\, a_{AI,t} = d) \propto \pi_H\left(d \,|\, \tilde{P}_{AI}(\boldsymbol{s})\right) \cdot P_H(\boldsymbol{s}) \;\; \forall \boldsymbol{s} \in \mathcal{S}.$$

Generative model of agent behavior Similar to prior work, we consider the human and the AI to behave as approximate planners (Wu et al., 2022, 2023), who tend to take the shortest path to the workplace. We assume that they choose a direction with a probability inversely proportional to $\text{ETA}(d \,|\, P_{agent})$, that is, the time they expect they will need to reach the workplace if their next movement is in direction d. To compute $\text{ETA}(d \,|\, P_{agent})$, we run Dijkstra's algorithm (Dijkstra, 1959) on a graph whose nodes correspond to tiles of the grid and edge weights represent the time required to move from one tile to the other averaged over states following from the agent's belief P_{agent}. Then, an agent's policy is given by the softmax

$$\pi_{agent}(d \,|\, P_{agent}) \propto e^{-\tau \cdot \text{ETA}(d \,|\, P_{agent})}. \tag{1}$$

Whenever the AI is in control, it selects a movement direction (*e.g.*, LEFT) and, with a probability p_{switch}, it may also ask the human for confirmation (*e.g.*, LEFT & ASK). If the human is in control, the AI decides between asking the human to switch or doing nothing, again with probability p_{switch}.

When the human encounters a prompt by the AI, they have to make a switching decision, that is, to decide whether they or the AI will drive the second half of the commute. We assume they behave rationally and they choose between the two

options proportionally to their probability of a successful outcome S. Let $P(S|P_H, \text{SWITCH})$, $P(S|P_H, \neg\text{SWITCH})$ be the success probability estimates of the human for each option. We assume that the human estimates these via Monte Carlo simulations. For the option that corresponds to them driving the second half, they perform L simulations of their driving behavior using Eq. 1 and compute the total success rate. For the option involving the AI, they sample L possible states $\boldsymbol{s} \sim P_H$ and, for each sample, they simulate the AI's driving using Eq. 1 and the belief $\tilde{P}_{AI}(\boldsymbol{s})$ introduced earlier. Based on the estimated probabilities of success, the human makes a (stochastic) decision $a_{sw} \in \{\text{SWITCH}, \neg\text{SWITCH}\}$ using the softmax

$$\pi_H(a_{sw}|P_H) \propto e^{\theta \cdot P(S|P_H, a_{sw})}. \quad (2)$$

Responsibility model

Given a commute instance generated by our environment, we predict responsibility judgments as a function of probabilities estimated by performing counterfactual simulations that use the aforementioned generative model. In our experiment, we focus on failure instances and thus, the counterfactual probabilities we consider here focus on counterfactual successes.

Human responsibility We predict that participants hold the human responsible for an observed failure relative to the extent that they would have succeeded had they made a different switching decision. Let a_{sw} denote the observed switching decision of the human. Then, we write the *counterfactual probability of success* as $P(S|a_{sw}, \texttt{do}(\neg a_{sw}))$, where $\texttt{do}(\cdot)$ denotes a counterfactual intervention (Pearl, 2009). Due to the multiplicity of counterfactual interventions in sequential decision-making (Tsirtsis & Gomez-Rodriguez, 2023; Tsirtsis et al., 2021) and the varying sensitivity of responsibility to each intervention's expectancy (Gerstenberg et al., 2018; Petrocelli et al., 2011), our model also considers the extent to which the alternative switching decision was expected. We will refer to this quantity as *counterfactual expectancy*, and we assume it is given by $\pi_H(\neg a_{sw}|P_H)$ and is proportional to the likelihood of success associated with the altenative decision (see Eq. 2). Our responsibility model considers the effects of the two factors both individually and jointly:

$$\begin{aligned} r_H = \alpha_1 &+ \alpha_2 \pi_H(\neg a_{sw}|P_H) + \alpha_3 P(S|a_{sw}, \texttt{do}(\neg a_{sw})) \\ &+ \alpha_4 \pi_H(\neg a_{sw}|P_H) \cdot P(S|a_{sw}, \texttt{do}(\neg a_{sw})) \end{aligned} \quad (3)$$

AI responsibility Our proposed model for the AI predicts that participants hold the AI responsible for an observed failure relative to the extent that the two agents would have succeeded if the AI had not assisted at all, and we write that counterfactual probability as $P(S|AI, \texttt{do}(\neg AI))$. Moreover, since the AI plays a more supportive role, we assume the participants' primary responsibility judgment is for the human, and the AI responsibility is complementary to the former. Let $\mathbb{1}[\text{AI}]$ denote the event that the AI drove for at least one tile. Then, our responsibility model takes the form

$$r_{AI} = \beta_1 + \beta_2 \mathbb{1}[\text{AI}] P(S|AI, \texttt{do}(\neg AI)) + \beta_3 (r_{max} - r_H). \quad (4)$$

Experiment

Our experiment asks participants to assign responsibility in a human-AI collaboration task (see Figure 1). We compare participants' responsibility judgments to the predictions of our responsibility model as well as a set of alternative models.

Methods

Participants The experiment was preregistered[2] and conducted online via Prolific. We recruited 50 participants (*age*: M = 37, SD = 12; *gender*: 31 female, 18 male, and 1 undisclosed; *race*: 5 Asian, 2 African American, 4 Multiracial, 38 White, and 1 undisclosed) who received $12/hour.

Procedure Participants were introduced to the semi-autonomous driving environment and the behavior of the two agents within it. They were asked 6 comprehension questions that they had to answer correctly before proceeding to the main experiment. The experiment consisted of 16 trials where the agents failed to reach the target destination on time.

On each trial, participants first watched an interactive step-by-step illustration of the respective commute, and then, they were asked to provide responsibility judgments while watching a video replay of the commute. The two questions ("to what extent is the [human / AI] responsible for not reaching on time?") were presented separately, and participants provided their responses with two continuous sliders ranging from 0 ("Not at all") to 100 ("Very much"). The average completion time of the experiment was 21 minutes (SD = 10).

Design The 16 trials of our experiment consist of 8 *twin trials*: pairs of trials where the observed commutes are exactly the same, but a small difference between the two grids alters the counterfactual outcome that would have occurred had the human made a different switching decision (see Fig. 2 for examples). To ensure participants do not recognize twin trials, we mirrored the twin gridworlds on the diagonal. The 8 twin trials manipulate 3 main factors: (i) whether the AI or the human is the initial driver, (ii) whether they switch control, and (iii) whether the decision of the human (not) to switch control was *right* or *wrong* at the moment that it was made. We will refer to that last factor as the human's *decision quality*, and we consider a decision to be right if the human believes that it leads to a higher probability of success (see Eq. 2). Across all trials, the path that each agent follows was sampled from our generative model given by Eq. 1. To manipulate factors (ii) and (iii), we generated switching decisions manually.

Results & Discussion

Do counterfactual outcomes influence human responsibility judgments? We investigate to what extent the way participants assign responsibility to the human differs depending on whether they would have reached the workplace on time had they made a different switching decision. To this end, we focus on pairs of twin trials and perform the following analysis. Let $r_H(p, tw[S])$ and $r_H(p, tw[F])$ denote the respon-

[2] https://osf.io/5ajzd

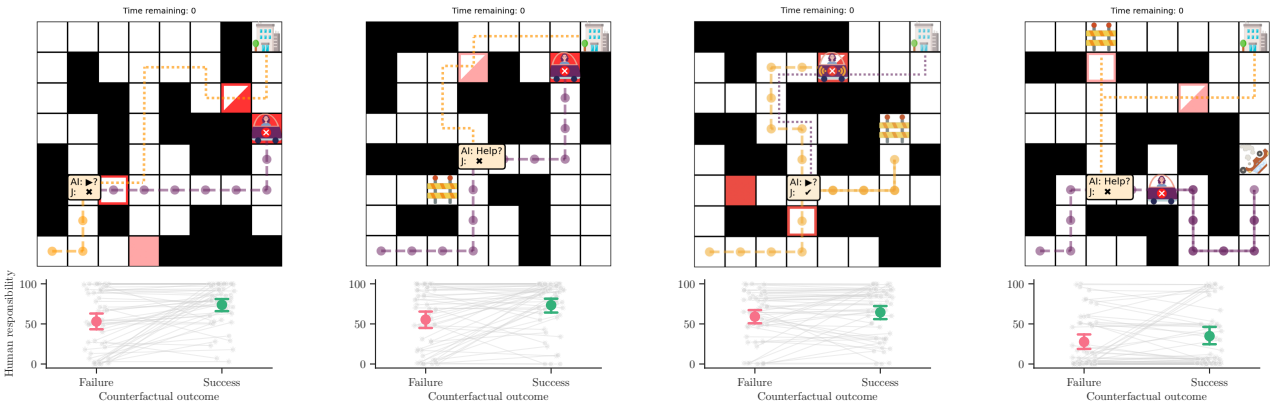
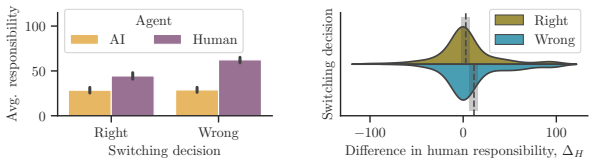


Figure 2: **Examples of twin trials and human responsibility judgments.** Each illustration shows a joint summary of two trials whose observed paths, outcomes, and decisions made by the agents are exactly the same. The grids of the two trials differ only in the congestion status of traffic spots illustrated as half colored (◪). In the trial where the traffic spot is not congested, had the human made a different switching decision, the agent who would have driven the second half would have reached the workplace on time following the dashed line. In the trial where the traffic spot is congested, the counterfactual outcome would have been a failure, same as the observed outcome. The figure below each illustration shows participants' judgments about the human's responsibility in the two twin trials. Colored points show means, and error bars show bootstrapped 95% confidence intervals. Each pair of gray points connected with a line shows the judgments of a single participant across the two twin trials.

sibility that a participant p assigns to the human in two twin trials with a counterfactual success (S) and failure (F), respectively. We denote as $\Delta_H(p,tw) = r_H(p,tw[S]) - r_H(p,tw[F])$ their difference. To quantify the effect of counterfactual outcomes on responsibility judgments, we fit a Bayesian linear mixed effects model with a fixed global intercept and random coefficients for each participant and pair of trials (*i.e.*, $\Delta_H \sim 1 + (1\,|\,p) + (1\,|\,tw)$). We observe that the global intercept's posterior mean is positive and equal to 6.48 (95% CI: $[-0.75, 13.78]$), which indicates that counterfactuals have a moderate effect on participants' judgments. To better understand this effect consider the examples in Figure 2. Many (but not all) participants hold the human more responsible for failing to reach on time whenever a different switching decision would have made a difference in the outcome. However, participants' judgments vary considerably, with some of them assigning equal or slightly less responsibility to the human.



(a) Average responsibility vs. decision quality

(b) Distribution of Δ_H vs. decision quality

Figure 3: **Effects of decision quality.** In panel (a), error bars indicate bootstrapped 95% confidence intervals. In panel (b), dashed lines show the means of the two distributions, and shaded areas illustrate 95% confidence intervals.

Does the human's decision quality make a difference to responsibility judgments? We first look at the average responsibility assigned to the human and the AI across trials where the human's switching decision is right and wrong, respectively. Figure 3a shows that the AI's average responsibility remains the same independently of the human's decision quality, while the human's responsibility increases when their decision was wrong. Moreover, across all trials, participants hold the human more responsible than the AI.

Additionally, we explore whether the effect of counterfactual outcomes on human responsibility judgments Δ_H varies depending on the quality of the switching decision. To test this, we use a dummy variable called *decision*, and set its value to 0 if the human's switching decision was right and 1 if it was wrong. We fit a Bayesian linear mixed effects model that includes an additional coefficient measuring the effect of the new variable (*i.e.*, $\Delta_H \sim 1 + decision + (1 + decision\,|\,p) + (1\,|\,tr)$). We observe that the mean for the posterior of the fixed coefficient of *decision* is positive and equal to 7.27 (95% CI: $[-5.67, 22.94]$). While its positive value indicates that participants may focus more on counterfactual outcomes whenever the observed switching decision was wrong, the effect is weak (the credible interval does not exclude 0). This can also be seen by looking directly at the distributions of Δ_H across pairs of twin trials with right and wrong decisions respectively (see Figure 3b). The two distributions are concentrated around zero but, in the case of wrong decisions, the distribution has a relatively larger mass on the positive side.

How well do the responsibility models capture participants' judgments? We start by estimating the required probabilities $\pi_H(\neg a_{sw}\,|\,P_H)$, $P(S\,|\,a_{sw}, \mathtt{do}(\neg a_{sw}))$ and

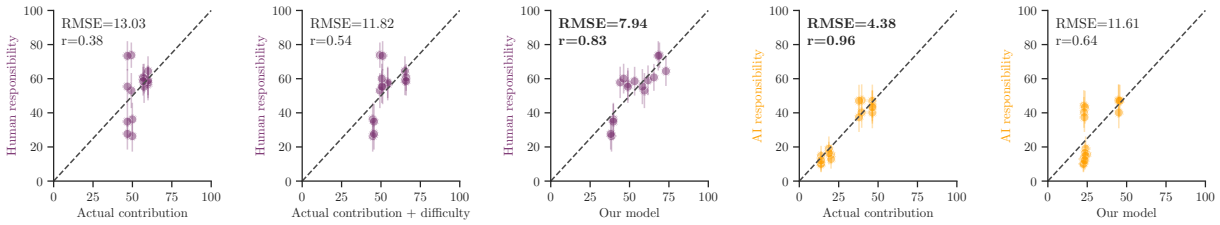


Figure 4: **Responsibility judgments and model predictions per trial.** Each point corresponds to one of our 16 trials, with the x-value showing the respective model prediction and the y-value showing the participants' average responsibility judgment. Different panels show results for the **human** and the **AI** under three models: (i) a simple model based on each agent's actual contribution to the outcome, (ii) an extension of the first model that also considers each trial's difficulty, and (iii) our proposed models given by Eqs. 3, 4. Across all panels, error bars indicate bootstrapped 95% confidence intervals.

$P(S \mid AI, \mathrm{do}(\neg AI))$ associated with each trial. We fix the hyperparameters τ and θ to the values 2 and 8 respectively and perform 300 Monte Carlo simulations in each grid. Then, we use the estimated probabilities along with participants' responsibility judgments to fit two Bayesian linear mixed effects models that take the form of Eqs. 3, 4 while also including random intercepts for individual participants. Additionally, we fit two baseline models. The first one assigns responsibility proportional to the respective agent's actual contribution to the outcome, measured as the number of time steps that the agent was in control of the car. For the human, we fit a model of the form $r_H \sim 1 + T_H + (1 \mid p)$, where T_H denotes the number of time steps that the human was in control and p denotes an individual participant. Similarly, for the AI, we fit a model that uses T_{AI}, the number of time steps that the AI was in control. The second baseline model is an extension of the first that includes the difficulty of the respective grid as an additional term, measured as the total number of obstacles (*i.e.*, road closures, traffic spots, and accidents).

To evaluate the different models, we first compare their average predictions per trial. Figure 4 shows the averaged model predictions per trial against participants' judgments. Our human responsibility model has the lowest RMSE and the highest correlation coefficient compared to the two baselines. In contrast, we observe that participants' judgments about the AI are best captured by the actual contribution model, although they didn't vary much across trials.

Table 1: **Model comparison.** Δelpd measures the predictive performance difference between each model and the best one. Lower values indicate worse performance. N-best shows the number of participants best captured by each model.

Model	Δelpd (se)	N-best
our model	0 (0)	3
additive effect	−2.4 (2.6)	7
counterfactual expectancy	−5.0 (3.6)	11
multiplicative effect	−27.5 (8.0)	5
actual contribution	−46.3 (11.1)	21
counterfactual prob. of success	−54.8 (10.5)	3

Because the models differ in their number of free parameters, we also compare them via approximate leave-one-out cross-validation (Vehtari et al., 2017) along with lesioned models that only contain individual components of our human responsibility model (*i.e.*, each additive term in Eq. 3). In total, we compare six models: (i) counterfactual expectancy, (ii) counterfactual probability of success, (iii) additive effect of (i, ii), (iv) multiplicative effect of (i, ii), (v) actual contribution, and (vi) our full model given by Eq. 3. Table 1 summarizes the results, which show that our model performs best overall. However, we observe that, when running cross-validations on individual participant responses, the actual contribution model best captures the most participants, followed by the model that uses counterfactual expectancy as predictor.

Conclusion

Although our responsibility model performed best overall, there were large individual differences (see Table 1). Those may arise from varying conceptions for how responsibility should be determined for human-AI collaborations and from participants' varying levels of motivation to carefully reason through the different scenarios (Lieder & Griffiths, 2020).

Our work opens up many interesting avenues for future work. Since the actual contribution model best captured the participants' judgments about the AI, it would be interesting to explore the relative importance of actual and counterfactual contribution, as well as how this mixture differs when making judgments about humans and AI agents (Xiang et al., 2023). In our setting, the AI and the human agent differ mainly in terms of what they know. It would be interesting to study settings where the agents differ in what they can do, too. To fit our responsibility model, we have set fixed values for the hyperparameters controlling the uncertainty of the model. In future work, it would be useful to conduct additional experiments to fit those hyperparameters by directly asking participants about counterfactual outcomes and the expectancy of the two agents' actions. Lastly, in our setting, the agents switch control at most once, and it would be interesting to explore situations that feature a more frequent back and forth between human and AI.

Acknowledgments

We thank Lara Kirfel for providing feedback on an early version of the manuscript and Sarah A. Wu, Sunny Yu and Nina Corvelo-Benz for their feedback on the experimental setup. We would also like to give credit to creators Freepik, Creartive, Smashicons, surang and juicy_fish from flaticon.com whose icons we have used to design our experiment. Tsirtsis and Gomez-Rodriguez acknowledge support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 945719). Gerstenberg acknowledges support from the Stanford Institute for Human-Centered Artificial Intelligence (HAI).

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological bulletin*, *126*(4), 556.

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2020). Drivers are blamed more than their automated cars when both make mistakes. *Nature human behaviour*, *4*(2), 134–143.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Balazadeh Meresht, V., De, A., Singla, A., & Gomez-Rodriguez, M. (2022). Learning to switch among agents in a team. *Transactions on Machine Learning Research*, *2022*(7), 1–30.

Beckers, S. (2023). Moral responsibility for AI systems. *Thirty-seventh Conference on Neural Information Processing Systems*.

Bernstein, D. S., Givan, R., Immerman, N., & Zilberstein, S. (2002). The complexity of decentralized control of markov decision processes. *Mathematics of operations research*, *27*(4), 819–840.

Byrne, R. M. (2016). Counterfactual thought. *Annual review of psychology*, *67*, 135–157.

Cañas, J. J. (2022). Ai and ethics when human beings collaborate with ai agents. *Frontiers in psychology*, *13*, 836650.

Chockler, H., & Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, *22*, 93–115.

De, A., Koley, P., Ganguly, N., & Gomez-Rodriguez, M. (2020). Regression under human assistance. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(03), 2611–2620.

De, A., Okati, N., Zarezade, A., & Gomez-Rodriguez, M. (2021). Classification under human assistance. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(7), 5905–5913.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, *1*(1), 269–271.

Franklin, M., Ashton, H., Awad, E., & Lagnado, D. (2022). Causal framework of artificial autonomous agent responsibility. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 276–284.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological review*, *128*(5), 936.

Gerstenberg, T., & Lagnado, D. A. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition*, *115*(1), 166–171.

Gerstenberg, T., & Tenenbaum, J. B. (2017, June). 515Intuitive Theories. In *The Oxford Handbook of Causal Reasoning*. Oxford University Press.

Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2018). Lucky or clever? from expectations to responsibility judgments. *Cognition*, *177*, 122–141.

Halpern, J., & Kleiman-Weiner, M. (2018). Towards formal definitions of blameworthiness, intention, and moral responsibility. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1).

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.

Kirfel, L., & Lagnado, D. (2021). Causal judgments about atypical actions are influenced by agents' epistemic states. *Cognition*, *212*, 104721.

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. *CogSci*.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive science*, *37*(6), 1036–1073.

Langenhoff, A. F., Wiegmann, A., Halpern, J. Y., Tenenbaum, J. B., & Gerstenberg, T. (2021). Predicting responsibility judgments from dispositional inferences and causal attributions. *Cognitive Psychology*, *129*, 101412.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, *43*, e1.

Lima, G., Grgić-Hlača, N., & Cha, M. (2021). Human perceptions on moral responsibility of ai: A case study in ai-assisted bail decision-making. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–17.

Longin, L., Bahrami, B., & Deroy, O. (2023). Intelligence brings responsibility-even smart ai assistants are held responsible. *Iscience*, *26*(8).

Mozannar, H., Bansal, G., Fourney, A., & Horvitz, E. (2022). Reading between the lines: Modeling user behav-

ior and costs in ai-assisted programming. *arXiv preprint arXiv:2210.14306*.

Okati, N., De, A., & Gomez-Rodriguez, M. (2021). Differentiable learning under triage. *Advances in Neural Information Processing Systems*, *34*, 9140–9151.

Oliehoek, F. A., Amato, C., et al. (2016). *A concise introduction to decentralized pomdps* (Vol. 1). Springer.

Pearl, J. (2009). *Causality*. Cambridge university press.

Petrocelli, J. V., Percy, E. J., Sherman, S. J., & Tormala, Z. L. (2011). Counterfactual potency. *Journal of personality and social psychology*, *100*(1), 30.

Raghu, M., Blumer, K., Corrado, G., Kleinberg, J., Obermeyer, Z., & Mullainathan, S. (2019). The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*.

Roese, N. J. (1997). Counterfactual thinking. *Psychological bulletin*, *121*(1), 133.

Shaver, K. G. (2012). *The attribution of blame: Causality, responsibility, and blameworthiness*. Springer Science & Business Media.

Straitouri, E., Singla, A., Meresht, V. B., & Gomez-Rodriguez, M. (2021). Reinforcement learning under algorithmic triage. *arXiv preprint arXiv:2109.11328*.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Triantafyllou, S., Singla, A., & Radanovic, G. (2022). Actual causality and responsibility attribution in decentralized partially observable markov decision processes. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 739–752.

Tsirtsis, S., De, A., & Gomez-Rodriguez, M. (2021). Counterfactual explanations in sequential decision making under uncertainty. *Advances in Neural Information Processing Systems*, *34*, 30127–30139.

Tsirtsis, S., & Gomez-Rodriguez, M. (2023). Finding counterfactually optimal action sequences in continuous state spaces. *Advances in Neural Information Processing Systems*, *36*.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*(9), 649–665.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, *27*, 1413–1432.

Wilder, B., Horvitz, E., & Kamar, E. (2021). Learning to complement humans. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*.

Wu, S. A., & Gerstenberg, T. (2024). If not me, then who? responsibility and replacement. *Cognition*, *242*, 105646.

Wu, S. A., Sridhar, S., & Gerstenberg, T. (2022). That was close! a counterfactual simulation model of causal judgments about decisions. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44).

Wu, S. A., Sridhar, S., & Gerstenberg, T. (2023). A computational model of responsibility judgments from counterfactual simulations and intention inferences. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

Xiang, Y., Landy, J., Cushman, F. A., Vélez, N., & Gershman, S. J. (2023). Actual and counterfactual effort contribute to responsibility attributions in collaborative tasks. *Cognition*, *241*, 105609.

Zultan, R., Gerstenberg, T., & Lagnado, D. A. (2012). Finding fault: Causality and counterfactuals in group attributions. *Cognition*, *125*(3), 429–440.