Responsibility in Multi-Agent Sequential Decision-Making: Comparing Human Judgments to Formal Models of Causal Attribution

Nripsuta Saxena[1*], **Stelios Triantafyllou**[2] and **Goran Radanović**[2]
[1]University of Southern California
[2]Max Planck Institute for Software Systems
nsaxena@usc.edu, strianta@mpi-sws.org, gradanovic@mpi-sws.org

Abstract

With the growing adoption of artificial intelligence in high-stakes decision-making, identifying the causes of outcomes–particularly failures–and determining who is responsible has become a critical concern. In this work, we examine how well formal definitions of *responsibility attribution*, grounded in the framework of *actual causality*, align with human judgments of responsibility. To this end, we conduct a large-scale survey to elicit human judgments of responsibility in multi-agent sequential decision-making scenarios, using a modified version of the card game Goofspiel. We evaluate multiple responsibility attribution methods, assess their alignment with human judgments about responsibility, and identify factors that significantly shape responsibility judgments. While no single responsibility attribution method consistently aligns with human responses, our findings highlight key factors that influence human responsibility judgments, including agent-specific biases and amount of information available to agents during decision-making.

1 Introduction

In high-stakes multi-agent sequential decision-making, attributing responsibility when harm occurs is a central concern for ensuring accountable outcomes. This task is inherently complex due to several intertwined challenges. Uncertainty about the consequences of individual and collective decisions makes it difficult to assess causal links between actions and outcomes. Furthermore, multi-agent settings face the *problem of many hands* [Van de Poel, 2015], where the actions of multiple agents interact in ways that blur the lines of individual responsibility. Temporal interdependence between decisions—where decisions made by one agent influence and are influenced by the actions of others over time—further complicates attribution. These factors imply that assigning responsibility cannot rely on simple causal or temporal proximity but instead requires careful modeling of agent interactions and their evolving contributions to outcomes.

Recent work attempts to tackle these challenges by providing a formal framework, grounded in actual causality, for attributing responsibility in complex multi-agent sequential decision-making environments [Triantafyllou *et al.*, 2022; Triantafyllou and Radanovic, 2023]. This framework treats the degree of responsibility as a quantitative measure of the extent to which an agent contributed to an outcome. The central idea is to identify the actual causes of the outcome, i.e., the specific decisions that were pivotal in bringing it about, and then determine an agent's responsibility based on whether their decisions appear among these causes and how significantly they contributed. This approach is axiomatic and prescriptive: it begins with formal principles that define actual causation and derives responsibility assignments by translating causal influence into measurable quantities.

However, these causal approaches to responsibility attribution do not incorporate human factors: how well they align with human judgments about responsibility is unclear. Yet, as argued by Lima et al. [Lima *et al.*, 2021; Lima *et al.*, 2023], public opinion matters, particularly when decisions have substantial societal impact. Hence, in this paper, we ask:

1. *How do responsibility attribution methods based on actual causality align with human judgments of responsibility?*
2. *Which factors most influence how humans assign responsibility in complex multi-agent scenarios?*

To this end, we design and conduct a human-subjects study ($n = 640$) to elicit human judgments about responsibility in multi-agent sequential decision-making[1]. We use a team-based version of the card game Goofspiel [Lanctot *et al.*, 2019; Triantafyllou *et al.*, 2022] as a representative setting for complex multi-agent scenarios. The game has two teams, Team A and Team B, each with one human and one AI player, making decisions over five rounds. Respondents are shown game scenarios—*vignettes*—in which Team A *loses* and are asked to assign a degree of responsibility to each player in Team A for the team losing.

We use multiple treatments to 1) test alignment of human judgments with responsibility attribution methods from the literature, and 2) quantify the influence of different factors that may affect human judgment. We study five responsibility attribution methods based on three definitions of actual

[*]Work performed during an internship at Max Planck Institute for Software Systems.

[1]The Ethical Review Board of the Faculty of Mathematics and Computer Science, Saarland University approved this study.

causality and two definitions of the degree of responsibility. We focus on three main factors that may influence judgments:

- **Initial condition**: Initial cards dealt to players were varied. In some games, Team A, or a member of Team A, was dealt worse cards than the other players, i.e., there was a *bias* against them. Our hypothesis was that people would assign a lower degree of responsibility to a player when they had a disadvantageous initial condition. Moreover, by using vignettes where only one of the players of Team A was subject to bias, we test if respondents distinguish between human and AI agents under unequal starting conditions.
- **State observability**: We ask respondents to assign responsibility assuming games were either *perfect-information* (i.e., players see each other's cards) or *imperfect-information* (i.e., players see only their own cards). Our hypothesis was that coordination under partial observability is more difficult, and thus, people would assign lower responsibility in those scenarios.
- **Aid in counterfactual reasoning**: As responsibility attribution methods are based on actual causality, counterfactual reasoning plays a critical role. We tested how judgments differ when counterfactuals are shown, i.e., when respondents are shown what would have happened if players from Team A had made different decisions. Our hypothesis was that counterfactuals help align human judgments more closely with attribution methods that rely on them.

Overview of results. Our findings show that human judgments of agent responsibility are context-sensitive. Greater information access and availability of counterfactuals increase perceived accountability: respondents assigned higher degrees of responsibility when players fully observed the state of the game and when counterfactuals were provided to the respondents. This suggests people calibrate their judgments based on the information available to the players and what they could plausibly do. Notably, under biased initial conditions, respondents assigned higher degrees of responsibility when the disadvantaged player was a human, but not when it was AI. This asymmetry suggests that people calibrate responsibility on the basis of a player's actions as well as the player's identity. Alignment with formal models of causal attribution was strongest under biased initial conditions, especially when both players in Team A were disadvantaged, suggesting that visible structural asymmetries prompt people to assess responsibility in ways that are more consistent with formal models. Finally, self-reported confidence was positively correlated with responsibility ratings and stronger alignment with formal models, suggesting that confidence reflects a sense of causal clarity.

Contributions. Our work contributes to the growing body of literature on understanding human judgments about responsibility in multi-agent settings involving humans and AI. Unlike prior work, we examine the alignment of formal approaches to responsibility attribution grounded in causality with human judgments, and we consider more complex environments, namely, sequential decision-making over multiple rounds. Our technical contributions are: 1) we introduce the first evaluation framework that bridges formal responsibility models with empirical human-judgment data, enabling direct and systematic comparison between responsibility assignments from formal models and human responsibility judgments; 2) we decompose key contextual features in multi-agent sequential decision-making settings, such as state observability, into independent, controllable treatments, allowing their effects on responsibility judgments to be studied systematically; and 3) we propose a new empirical framework that uses visual, sequential scenarios to make concepts related to actual causality (AC) and degree of responsibility (DR) understandable to laypeople. Our modular pipeline enables any DR method grounded in AC to be evaluated in the future, as well as the addition of new contextual treatments.

2 Related Work

Actual causality. This paper relates to an extensive body of literature on actual causality. Much prior work on actual causality has focused on extending the But-For definition–henceforth, the BF definition–which is based on the *but-for* test [Hart and Honoré, 1985; Pearl, 1998; Hitchcock, 2001; Hall, 2007; Halpern and Hitchcock, 2015]. Under this criterion, a set of agents' actions is considered a cause of an outcome if the outcome would not have materialized had this set of actions been altered. One extension of the BF definition was proposed by Halpern and Pearl (2015), henceforth, the HP definition. The HP definition extends the classical BF definition by introducing contingencies. We refer the reader to the related work section of Triantafyllou (2022), which provides a brief overview of these definitions and explains how they relate to the HP definition and its variants [Halpern and Pearl, 2005; Halpern, 2015; Halpern, 2016]. Our goal is to study human perceptions in complex multi-agent sequential decision-making. Hence, we focus on definitions that have been operationalized in one such setting (i.e., in Dec-POMDPs) [Triantafyllou *et al.*, 2022]. In contrast to past work that studies the computational and structural properties of these definitions [Triantafyllou and Radanovic, 2023], we examine how responsibility-attribution methods based on these definitions align with human responsibility judgments and the factors that contribute to this alignment.

Formal models of responsibility. The most relevant prior work to ours concerns models of responsibility based on actual causality [Chockler and Halpern, 2004; Triantafyllou *et al.*, 2022; Triantafyllou and Radanovic, 2023; Halpern and Kleiman-Weiner, 2018; Alechina *et al.*, 2020], among which we focus on those operationalized in Dec-POMDPs [Chockler and Halpern, 2004; Tigard, 2021; Triantafyllou and Radanovic, 2023]. However, we note that prior research has also considered alternative approaches to responsibility attribution. Baier et al. (2021) propose a game-theoretic framework for responsibility attribution, distinguishing between two notions of responsibility: forward (prescriptive) and backward (retrospective). The latter is related to causal notions of responsibility [Chockler and Halpern, 2004] and attributes responsibility for a realized play, whereas the former attributes responsibility based on all potential plays. Hence, it is similar to the notion of blame in multi-agent Markov Decision Processes proposed by Triantafyllou et al. (2021), which considers average performance rather than realized outcomes

when attributing blame. Other works on causal responsibility distinguish between responsibility, blame, and blameworthiness [Chockler and Halpern, 2004; Halpern and Kleiman-Weiner, 2018], incorporating agents' epistemic states and intentions into formal models. Much prior work quantifies responsibility (or related concepts) using concepts from cooperative game theory, such as the Shapley value [Shapley, 2016; Shapley and Shubik, 1954] and Banzhaf index [Banzhaf III, 1964; Banzhaf III, 1968]. Other approaches explore computational models of responsibility judgments based on counterfactual simulation [Tsirtsis *et al.*, 2024]. Our contribution to this line of work is to understand the alignment between formal models of responsibility based on actual causality and human responsibility judgments.

Moral responsibility. Work on morality goes back centuries with philosophers such as David Hume and Jeremy Bentham offering early systematic accounts of moral judgments in the mid-18th century [Hume, 2023; Morris and Brown, 2001]. Historically, normative philosophy centered on the principle of harm avoidance–the idea that moral action requires avoiding and preventing harm, often treating other moral considerations, such as honesty, as secondary when in conflict with this imperative [Mill, 2001; Bentham, 1988].

Along those lines, a fundamental aspect of morality involves making judgments about when individuals are morally responsible for their actions and holding them accountable for the resulting consequences. While an agent's moral responsibility may differ from their causal responsibility (for example, a toddler causing an adverse outcome may be causally but not morally responsible), the two are typically closely intertwined [Talbert, 2025]. This interplay between causal and moral responsibility becomes especially salient in an era of increasing human-machine interaction. As autonomous systems take on greater roles in decisions that may affect people's lives, understanding how society perceives responsibility for adverse outcomes has become increasingly vital. Although responsibility and blame are distinct, with *interpersonal blame* philosophically considered a response to moral agents on the basis of their adverse actions [Tognazzini and Coates, 2025], research on how people assign blame offers valuable insight into broader perceptions of responsibility.

Human attitudes and perceptions. When harm is caused by only one agent, prior work shows people blame machines more than humans for the same mistake [Hong *et al.*, 2020; Franklin *et al.*, 2021]. Similarly, robots are blamed more than humans when they fail to make utilitarian decisions [Malle *et al.*, 2014; Lima *et al.*, 2021]. This trend reverses in settings involving human-robot interaction under human supervision (e.g., a semi-autonomous vehicle operating under the supervision of a human driver), where people blame the human more than the machine [Awad *et al.*, 2020; Beckers *et al.*, 2022]. Autonomous systems are judged more harshly than non-autonomous agents that rely on human input [Furlough *et al.*, 2021; Kim and Hinds, 2006]. In both cases, the human was blamed more than the agent. Recent work also studies how people attribute blame and causal responsibility in settings where human and AI agents act independently yet contribute jointly to an adverse outcome: in Meier et al. (2025), when two agents perform identical actions but only one violates an expected norm, the norm-violating agent is consistently judged more blameworthy and causally responsible, regardless of whether it was human or a machine.

We extend this literature by examining whether formal definitions of responsibility attribution align with human judgments in multi-agent, sequential decision-making settings where humans collaborate with AI agents. We study a more complex setting than prior work: humans and AI agents make sequential choices, rendering the attribution of responsibility for an adverse outcome substantially more intricate. Moreover, our work is the first to study how formal causality-based responsibility models align with human judgments.

3 Methodology

Our survey was based on a version of the game *Goofspiel* [Triantafyllou *et al.*, 2022; Lanctot *et al.*, 2019; Meng *et al.*, 2023], which we used to simulate interactions between two competing teams, each comprising one human and one AI player. We conducted a survey experiment with respondents recruited via the Prolific platform. Respondents were asked to evaluate agent responsibility across a range of carefully designed Goofspiel vignettes. We first describe causal responsibility attribution briefly, followed by an introduction to *Goofspiel* and the vignettes shown to respondents. We then detail the treatments in the survey and our recruitment protocol.[2]

3.1 Responsibility Attribution Methods

We study responsibility attribution (RA) methods that first identify actual causes of an outcome using one of the definitions of actual causality from the AI literature [Halpern, 2016; Triantafyllou *et al.*, 2022]. We consider three definitions of actual causation (AC): in addition to the BF definition [Hart and Honoré, 1985] and the HP definition [Halpern and Pearl, 2005], we consider a definition introduced in Triantafyllou et al. (2022), henceforth, the TR definition. Once the actual causes of an outcome have been identified, the question becomes: how much responsibility should each agent bear for that outcome? To answer this, we can define the *degree of an agent's responsibility*. We consider two definitions of the degree of responsibility (DR): one by Chockler and Halpern (2004) (henceforth, the CH degree of responsibility) and another by Triantafyllou et al. (2022) (henceforth, the TR degree of responsibility). Note that the CH and TR degrees of responsibility are agnostic to the specific definition of actual causality—any definition can be used to determine actual causes. This gives six possible AC$\times$DR combinations. However, two combinations (BF$\times$CH and BF$\times$TR) yield the same degrees of responsibility, thus, we have five distinct responsibility attribution methods. We consider these definitions within the decentralized partially observable Markov decision process (Dec-POMDP) framework. We refer the reader to Triantafyllou et al. (2022) for more details.

3.2 The Goofspiel game

We consider a team-based variant of the card game *Goofspiel*, as in Triantafyllou et al. (2022), with five rounds. Two

[2] An extended version of the paper is available on arXiv with additional implementation and analysis details (e.g., regression tables).

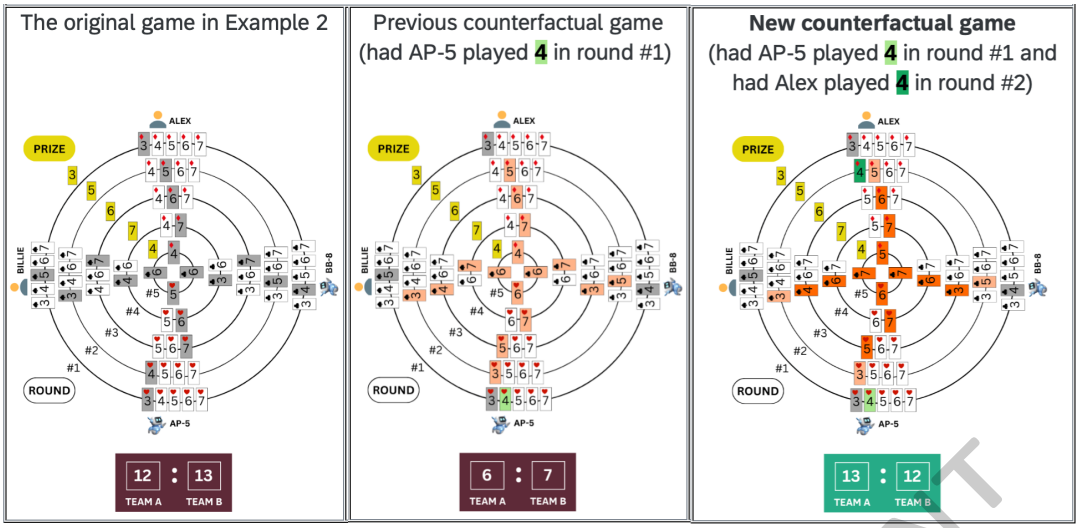

Figure 1: An example Goofspiel match with counterfactuals. The left panel shows the original game (gray cards); the middle and right panels show counterfactual games with first and second interventions highlighted in light green and dark green respectively, and subsequent actions shown in light and dark orange, respectively.

teams – Team A and Team B – compete, each with one human player and one AI agent (Figure 1). Team A includes Alex (human) and AP-5 (AI); Team B has Billie (human) and BB-8 (AI). The human players have gender-neutral names [Brylla, 2009], and the AI players' names were inspired by droids from *Star Wars* [Disney, a; Disney, b].

We use "players" and "agents" interchangeably throughout the paper. Figure 1 depicts our game setup: the original game (left panel) and two counterfactuals (middle, right panels).

Description of Goofspiel. At the start of the game, each player is dealt five cards from the standard 52-card deck. All players are aware that any given player holds cards of only one suit. Additionally, a separate deck of cards—called prize deck—is shuffled and placed face down. The game proceeds over five rounds. In each round, a prize card is drawn from the prize deck and revealed to all players; its value is shown in yellow between the players Alex and Billie. Upon observing this prize card, each player simultaneously selects one card from their hand to play. The selected card is shown in color, while the remaining unplayed cards appear in white. The team whose combined bid (i.e., the sum of the two selected cards) is higher wins the round and receives the prize card, which is added to the team's score. If the teams have equal sums, the prize card is discarded and no points are awarded (i.e., the scores remain the same). All cards played in that round are discarded, and a new round begins. The game ends after round 5, and the team with the highest cumulative score wins. If the teams have equal scores, then the game is a draw.

Counterfactual Games. The left panel of Figure 1 depicts the actual sequence of play, while the middle and right panels present counterfactual versions of the game. In the counterfactual in the middle panel, AP-5 selects a different card in Round 1. This alternative card AP-5 plays in the counterfactual is highlighted in light green, while the card originally played by AP-5 in that round is shown in gray. Following the literature on causality, we refer to this as an *intervention* [Halpern and Pearl, 2005; Triantafyllou *et al.*, 2022]. This alters the composition of AP-5's hand, influencing subsequent rounds, where AP-5 may sometimes choose the same card as in the original sequence and at other times a different one. Cards played after the initial intervention are shown in light orange, signaling both the counterfactual nature of the game and the intervention itself. At its core, a counterfactual game poses the question: *What would have happened if AP-5 had played card #4 in Round 1?*

Counterfactual games may include multiple interventions. Consider the right panel of Figure 1. AP-5 plays a different card in Round 1, while Alex chooses a different card in Round 2. This scenario effectively asks: *What would have happened if AP-5 had played card #4 in Round 1 and Alex had played card #4 in Round 2?* In such cases, the first intervention is highlighted in light green and the second in dark green. Cards played after the first intervention are shown in light orange, while those played after the second intervention are shown in dark green. As with simpler counterfactuals, in later rounds players may either replicate the card they played in the original game or select different cards. When both interventions occur in the same round, they are jointly highlighted in light green, and all subsequent plays are marked in light orange.

Intervention(s) in counterfactual games can lead to a decisive shift in outcome. Comparing the original game and its counterfactuals shows how different decisions by one or both agents can causally alter the game's final result: in Figure 1 (right panel), the result is reversed–Team A wins instead.

Game Scenario Generation. We generate Goofspiel games by rolling out policies from Triantafyllou et al. (2022). For Team A, we use their policy of agent Ag0, while for Team B, we use their stochastic Opponents' policies. The initial configuration of each game depends on the *initial condition* treatment (see Sec 3.4). The generation process otherwise follows that of Triantafyllou et al. (2022): (1) before the game starts we shuffle the prize deck; (2) in each round we sample

Team B's actions. With this process, we obtain 1000 games per initial configuration in which Team A did not win.

Like Triantafyllou et al. (2022), we further identify actual causes and degrees of responsibility for each of the original games under each combination of actual causality definitions and responsibility attribution methods. In contrast to Triantafyllou et al. (2022), we limit the number of interventions in counterfactual games to 2 instead of 4. In treatments with counterfactual scenarios, the set of actual causes for a given game defines the counterfactual scenarios shown to respondents. For example, in a treatment where we present one counterfactual scenario per game relevant to calculating the degree of responsibility of AP-5, we use the actual cause most relevant for determining AP-5's degree of responsibility. In a treatment where we present two counterfactual scenarios relevant to calculating the degrees of responsibility of AP-5 and Alex, we analogously use the actual causes most relevant for determining their respective degrees of responsibility. Tutorials for some treatments include counterfactual games that are not based on actual causes but serve as illustrative examples for explaining counterfactuals (middle panel of Figure 1). Counterfactual games generated by actual causes change the outcome of the original game (right panel of Figure 1).

3.3 Survey Design: Vignettes

The respondents first complete a tutorial explaining the rules of *Goofspiel*, counterfactual games, and the evaluation task to be performed. Respondents were not given a description of the game scenario generation process. After the tutorial, each respondent viewed five Goofspiel game vignettes in which Team A loses. Each vignette included none, one, or two different counterfactual games, depending on the aid in counterfactual reasoning treatment (Sec 3.4). These counterfactual games, generated using actual causes as explained above, lead to Team A winning. For each vignette, the respondent rated the degree of responsibility of each Team A player (Alex and AP-5) for Team A losing the original game on a 3-point Likert scale (*Low*, *Medium*, or *High Responsibility*), and also provided confidence in their judgments on a second 3-point Likert scale (*Low*, *Medium*, or *High Confidence*).

3.4 Survey Design: Treatments

To study factors that shape human judgments of responsibility, we vary four key features of the decision-making context:

1. **Responsibility attribution method**, i.e., the definitions of actual causality and degree of responsibility used to generate counterfactual scenarios for respondents and reference values for degrees of responsibility of Alex and AP-5;
2. **Initial condition**, i.e., whether there is any bias in the initial hands of different players;
3. **State observability**, i.e., the information available to players during the game;
4. **Aid in counterfactual reasoning**, i.e., the amount of counterfactual information shown to respondents.

We use between- and within-subjects manipulations across these dimensions to isolate their individual and combined effects on responsibility judgments. The first treatment is within-subjects, the rest are between-subjects (see Table S2).

Responsibility attribution method (AC×DR). This treatment varies the logic used to generate counterfactual games and reference degrees of responsibility, using five distinct responsibility attribution methods, i.e., AC×DR combinations, detailed in Section 3.1. This treatment enables us to study the alignment between responsibility attribution methods based on actual causality and human judgments about responsibility. Respondents viewed one scenario per AC×DR combination, enabling a within-subjects comparison across possible AC×DR combinations. The order of the combinations was randomized for each participant.

Initial condition. This treatment introduces asymmetries in strategic advantage by manipulating the initial distribution of card values across players, simulating real-world inequities in resource access and opportunity. Over four conditions, players begin with stronger or weaker hands, affecting their ability to act effectively. In the *unbiased* condition, all players receive the same cards (of value 3-7), ensuring a level playing field. The three *biased* conditions introduce targeted disadvantages. In *bias against Alex*, Alex receives low-value cards (2–6), while other players receive cards (3–7). In *bias against AP-5*, AP-5 is disadvantaged and receives low-value cards (2–6), while other players receive cards (3–7). In *bias against both*, both Alex and AP-5 receive low-value cards (2–6), while their opponents receive the default cards (3–7).

This treatment enables us to study how biased resource allocations shape human responsibility judgments in multi-agent sequential decision-making, particularly in mixed human–AI teams, and whether respondents distinguish between human and AI agents under unequal starting conditions. Respondents were randomly assigned to one condition.

State observability. This treatment manipulates the perceived information structure of the game by varying what respondents are told about the information available to the players. There are two conditions: *imperfect-information* and *perfect-information* games, which depict common settings in multi-agent systems [Shoham and Leyton-Brown, 2008].

In the *imperfect-information* condition, respondents are told players can see only their own hand, implying they must decide which card to play without knowing the other players' cards. In the *perfect-information* condition, respondents are told all players see every player's hand, implying players are more certain about the state of the game. A note in each vignette explicitly states the information structure.

This treatment offers a natural testbed to examine how the amount of information the players had access to during the game affects people's perceptions of responsibility in human-AI teams. It enables us to study if the uncertainty introduced due to imperfect information affects responsibility attribution. Respondents were randomly assigned to one condition.

Aid in counterfactual reasoning. This treatment varies the amount of counterfactual information shown to respondents. Respondents were randomly assigned to one of the four conditions below. In the *Factual-only* condition, respondents see only the original game, i.e., the players' original actions and the outcome (Team A losing). In other conditions, one or more counterfactuals are shown with the original:

- *Counterfactual for Alex*: respondents additionally see a

counterfactual game generated from the actual cause that is most relevant for determining the degree of Alex's responsibility using the underlying AC × DR condition;

- *Counterfactual for AP-5*: respondents additionally see a counterfactual game generated from the actual cause that is most relevant for determining the degree of AP-5's responsibility using the underlying AC × DR condition;
- *Counterfactuals for both*: respondents additionally see two counterfactual games from the previous two conditions (*counterfactual for Alex* and *for AP-5*).

Varying the counterfactual(s) shown (Alex's, AP-5's, both, or none) tests if judgments shift from what *did* occur to what *could have occurred*, based on each agent's perceived capacity to act otherwise. Respondents were randomly assigned to one treatment. See Figures S4-S6 in supplementary materials for vignettes of different treatment conditions.

This treatment studies how explicit counterfactuals influence perceptions of responsibility, probing two key questions:

1. How sensitive are humans' responsibility judgments to the presence and extent of counterfactual comparisons?
2. Do these effects differ based on whether the counterfactual of the agent in question is human or AI?

3.5 Survey Design: Recruitment Protocol

An a priori power analysis determined the sample size to detect medium effects (Cohen's d $= 0.5$) with 80% power at $\alpha = 0.05$. The AC×DR treatment's within-subjects design (five vignettes per participant) increased statistical power by reducing individual variability. Between-subjects treatments needed 20 respondents per condition, yielding a target sample of 640 across 32 treatment combinations. Respondents were recruited on Prolific: U.S.-based users with $> 95\%$ approval and ≥ 100 completed tasks for data quality. Each was randomly assigned to a treatment combination and paid US$3.41 (over U.S. federal minimum wage).

4 Results

We discuss how each treatment affected responsibility ratings for Alex and AP-5, followed by the alignment of the respondents' judgments with formal AC×DR assignments.

4.1 Human Responsibility Judgments

We fit a multi-variate linear mixed-effects regression model designed to predict respondents' responsibility ratings for the human (Alex) and AI (AP-5) players, allowing us to test whether the experimental treatments systematically influenced the degree of responsibility respondents for each agent. The model includes fixed effects for all key treatments (ACxDR combination, initial condition, state observability, aid in counterfactual reasoning), and self-reported confidence. To account for repeated-measures–each respondent rated five distinct Goofspiel vignettes–we include random intercepts for respondent ID and vignette ID, capturing individual- and scenario-level variability. All $p-$values reflect tests using the standard significance threshold, $\alpha = 0.05$.

Effect of actual causality and responsibility definitions. We evaluate how different AC×DR combinations affect respondents' responsibility ratings; as noted earlier, AC×DR determine counterfactuals shown to respondents. While no AC×DR combination showed a statistically significant main effect, some (e.g., HP×TR, $p \approx 0.09$) exhibited marginal trends. This suggests that certain formalisms may exert a greater effect on human judgments than others, implicitly through the choice of counterfactuals shown to human raters.

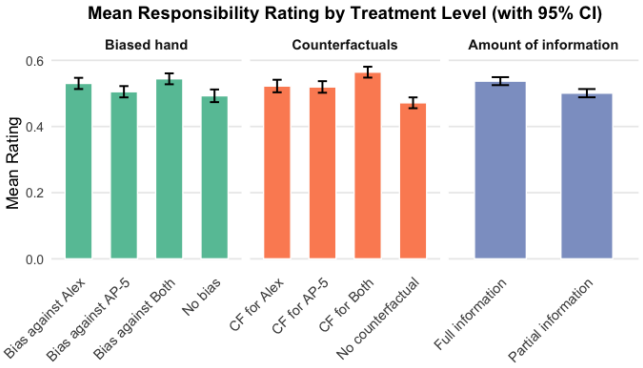

Figure 2: Mean responsibility rating. Colors represent treatment types; bars within each color indicate levels of that treatment.

Effect of biased initial condition. Bias in initial hands produced a nuanced but noteworthy effect on responsibility attributions. Responsibility ratings increased significantly when the *human player, Alex, was disadvantaged* by holding worse cards ($p = 0.0182$) and when *both the human player, Alex, and the AI player, AP-5, were disadvantaged* with worse cards ($p = 0.0111$), relative to the unbiased baseline in which all players began with identical cards (left plot in Figure 2). By contrast, *bias against AI player, AP-5, alone* did not yield a significant shift from the baseline ($p = 0.3257$).

These findings reveal an asymmetry in how biased resource allocation affects responsibility attribution: human players (Alex) are held more accountable under disadvantaged conditions than their AI teammate. This suggests responsibility judgments are sensitive to fairness in resource distribution and who the disadvantaged party is (human or AI). This is consistent with past research that perceptions of fairness, agency, and obligation are differentially applied to human and AI agents [Awad *et al.*, 2020; Lima *et al.*, 2023].

Effect of information available to players. Next, we evaluate how state observability shapes responsibility judgments. Respondents assigned significantly greater responsibility in the full-information condition than in the partial-information condition ($p = 0.025$) (Fig 2, right plot). These findings indicate responsibility judgments are sensitive to the amount of information available to players at the time of decision-making, with access to more information leading to harsher judgments of responsibility. This suggests laypeople expect heightened accountability when agents are better informed and better positioned to act effectively, consistent with prior research linking transparency and access to more information to increased perceptions of accountability[Awad *et al.*, 2020].

Effect of number of counterfactuals shown. Showing counterfactuals significantly increased responsibility ratings, with the strongest effect when both agents' counterfactuals were presented ($p < 0.001$). Compared to the *factual-only*

baseline, responsibility ratings increased with a *counterfactual for Alex* ($p = 0.0312$) or *counterfactual for AP-5* ($p = 0.0126$), and were highest when *counterfactuals for both* were shown($p < 0.001$) (middle plot, Fig. 2). This demonstrates counterfactual framing strongly shapes perceived responsibility, underscoring the importance of how explanations are framed in human–AI systems. In other words, what is shown or omitted can systematically bias responsibility attributions, highlighting the need for careful consideration of how hypothetical alternatives or explanations are communicated, especially when accountability is critical.

Effect of confidence. We observe a strong positive correlation between self-reported confidence and responsibility ratings ($p < 0.001$): respondents with greater certainty in their judgments were more likely to assign higher levels of responsibility. This suggests subjective confidence may serve as a useful proxy for perceived causal clarity—when individuals feel more confident they may perceive the causal structure of the scenario as more determinate or interpretable.

4.2 Alignment with Formal Responsibility Models

We study the extent to which respondents' responsibility judgments align with the responsibility assignments produced by different AC×DR combinations. For each respondent-vignette pair, we compute an *agreement score* quantifying how closely respondents' ratings of Alex and AP-5 match the responsibility levels prescribed by the AC×DR combination used in the vignette. Agreement scores range from 0 (no alignment) to 1 (full alignment), with an intermediate score of 0.5 indicating agreement for one player but not the other.

To capture how well respondents' judgments reflect the AC×DR combinations presented in the vignettes, we estimate a multivariate linear mixed-effects regression model with the agreement score as the dependent variable. Fixed effects include experimental treatments (ACxDR combination, initial condition, state observability, aid in counterfactual reasoning), and respondents' self-reported confidence. Random intercepts are included for respondent ID and vignette ID to account for repeated measures and scenario-specific variation. This framework allows us to isolate the impact of each treatment on alignment with formal models of responsibility attribution while appropriately modeling the nested structure of the data. (Regression tables in supplementary materials.)

Effect of biased initial condition. We find that biased hands drive alignment. Respondents in the *bias against both agents* condition showed significantly higher agreement with the AC×DR treatment-prescribed responsibility attributions ($p = 0.0027$), while those in the *unbiased* condition were significantly less aligned ($p = 0.0033$). This suggests people not only take structural disadvantage into account when assessing responsibility, but are also more aligned with formal models of responsibility attribution when visible bias is present. In contrast, when no visible bias is present, respondents may rely more on intuitive or heuristic judgments.

Effects of AC×DR combination, available information, and number of counterfactuals presented. We found no significant effects on respondents' alignment with the AC×DR treatment-prescribed responsibility assignments across these three treatments– AC×DR combination, state observability, aid in counterfactual reasoning. Limiting agents' access to information did not significantly affect agreement, nor did the selective presentation of counterfactuals meaningfully shift agreement. Finally, we found no significant differences in agreement scores across AC×DR combinations–that is, no particular AC×DR combination consistently aligned more closely with respondents' responsibility ratings for Alex and AP-5. This suggests respondents did not systematically favor one responsibility attribution method over another, underscoring the difficulty of capturing human moral intuitions through formal causal frameworks. This highlights the need for further research to understand how structural, informational, and conceptual factors interact to shape human agreement with formal causal frameworks in collaborative human-AI decision-making contexts.

Effect of confidence. Self-reported confidence was also a significant predictor of agreement ($p = 0.0015$). More confident respondents were more likely to align with the AC×DR treatment-prescribed responsibility assignment, further supporting the idea that confidence reflects a sense of causal clarity or coherence with one's internal model of responsibility.

5 Discussion

This work presents the first independent evaluation framework for assessing degree-of-responsibility (DR) models grounded in actual causality (AC). Using this framework, we systematically study how human perceptions of responsibility align with formal responsibility attribution methods grounded in actual causality. Through a large-scale survey based on a team-based version of Goofspiel, we examined how people assign responsibility to human and AI agents in collaborative, sequential decision-making settings. We found a biased initial hand selectively increases perceived responsibility for the human agent, Alex, but not for the AI agent, AP-5; full-information settings yield higher responsibility ratings; and the presence of counterfactuals substantially amplified perceived responsibility, especially when counterfactuals are shown for both agents. Alignment between human judgments and formal responsibility attribution models was strongest when structural asymmetries were present in the players' initial hands, although no responsibility attribution model significantly matched human judgment across all contexts.

Our results reveal important gaps between current formal models of responsibility attribution and human judgments. In particular, our findings highlight that existing models do not capture several factors that humans consider when assigning responsibility, including initial conditions, information availability, and agent type (human vs. AI). These observations suggest the need for more sophisticated formal models of responsibility attribution, and present a promising direction for future research. Finally, while we chose Goofspiel because it is a standard benchmark in multi-agent reinforcement learning and enables controlled manipulation of contextual features that are general across multi-agent systems (e.g., initial conditions), we acknowledge that it may not capture all aspects of a multi-agent interaction. Therefore, exploring orthogonal aspects, including those which may be domain-specific, also represent an important avenue for future work.

Acknowledgments

This research was, in part, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 467367360.
We thank Nina Grgić-Hlača for insightful feedback and constructive discussions.

References

[Alechina *et al.*, 2020] Natasha Alechina, Joseph Y. Halpern, and Brian Logan. Causality, responsibility and blame in team plans. *arXiv preprint arXiv:2005.10297*, 2020.

[Awad *et al.*, 2020] Edmond Awad, Sydney Levine, Max Kleiman-Weiner, Sohan Dsouza, Joshua B Tenenbaum, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. Drivers are blamed more than their automated cars when both make mistakes. *Nature human behaviour*, 4(2):134–143, 2020.

[Baier *et al.*, 2021] Christel Baier, Florian Funke, and Rupak Majumdar. A game-theoretic account of responsibility allocation. *arXiv preprint arXiv:2105.09129*, 2021.

[Banzhaf III, 1964] John F. Banzhaf III. Weighted voting doesn't work: A mathematical analysis. *Rutgers Law Review*, 19:317, 1964.

[Banzhaf III, 1968] John F. Banzhaf III. One man, 3.312 votes: a mathematical analysis of the electoral college. *Villanova Law Review*, 13:304, 1968.

[Beckers *et al.*, 2022] Niek Beckers, Luciano Cavalcante Siebert, Merijn Bruijnes, Catholijn Jonker, and David Abbink. Drivers of partially automated vehicles are blamed for crashes that they cannot reasonably avoid. *Scientific reports*, 12(1):16193, 2022.

[Bentham, 1988] Jeremy Bentham. The principles of morals and legislation, 1988.

[Brylla, 2009] Eva Brylla. Female names and male names. equality between the sexes. 2009.

[Chockler and Halpern, 2004] Hana Chockler and Joseph Y Halpern. Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, 22:93–115, 2004.

[Disney, a] Star Wars Disney. List of star wars characters: Ap-5.

[Disney, b] Star Wars Disney. List of star wars characters: Bb-8.

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

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

[Hall, 2007] Ned Hall. Structural equations and causation. *Philosophical Studies*, 132(1):109–136, 2007.

[Halpern and Hitchcock, 2015] Joseph Y. Halpern and Christopher Hitchcock. Graded causation and defaults. *The British Journal for the Philosophy of Science*, 66(2):413–457, 2015.

[Halpern and Kleiman-Weiner, 2018] Joseph Y. Halpern and Max Kleiman-Weiner. Towards formal definitions of blameworthiness, intention, and moral responsibility. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[Halpern and Pearl, 2005] Joseph Y Halpern and Judea Pearl. Causes and explanations: A structural-model approach. part i: Causes. *The British journal for the philosophy of science*, 2005.

[Halpern, 2015] Joseph Y Halpern. A modification of the halpern-pearl definition of causality. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 3022–3033, 2015.

[Halpern, 2016] Joseph Y Halpern. *Actual causality*. MiT Press, 2016.

[Hart and Honoré, 1985] Herbert Lionel Adolphus Hart and Tony Honoré. *Causation in the Law*. OUP Oxford, 1985.

[Hitchcock, 2001] Christopher Hitchcock. The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy*, 98(6):273–299, 2001.

[Hong *et al.*, 2020] Joo-Wha Hong, Yunwen Wang, and Paulina Lanz. 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, 2020.

[Hume, 2023] David Hume. *A treatise of human nature: Being an attempt to introduce the experimental method of reasoning into moral subjects*. Broadview Press, 2023.

[Kim and Hinds, 2006] Taemie Kim and Pamela Hinds. Who should i blame? effects of autonomy and transparency on attributions in human-robot interaction. In *ROMAN 2006-The 15th IEEE international symposium on robot and human interactive communication*, pages 80–85. IEEE, 2006.

[Lanctot *et al.*, 2019] Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, et al. Openspiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*, 2019.

[Lima *et al.*, 2021] Gabriel Lima, Meeyoung Cha, Chihyung Jeon, and Kyung Sin Park. The conflict between people's urge to punish ai and legal systems. *Frontiers in Robotics and AI*, 8:756242, 2021.

[Lima *et al.*, 2023] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. Blaming humans and machines: What shapes people's reactions to algorithmic harm. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–26, 2023.

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

[Meier *et al.*, 2025] Jeremy Meier, Richard Wylie, and Simon M Laham. Who's to blame? blame attribution in autonomous human-machine interactions. *Blame Attribution in Autonomous Human-Machine Interactions (June 16, 2025)*, 2025.

[Meng *et al.*, 2023] Linjian Meng, Zhenxing Ge, Pinzhuo Tian, Bo An, and Yang Gao. An efficient deep reinforcement learning algorithm for solving imperfect information extensive-form games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5823–5831, 2023.

[Mill, 2001] John Stuart Mill. Utilitarianism. edited by george sher. *Indianapolis: Hackett*, 2001.

[Morris and Brown, 2001] William Edward Morris and Charlotte R Brown. David hume. *Stanford Encyclopedia of Philosophy*, 2001.

[Pearl, 1998] Judea Pearl. On the definition of actual cause. Technical report r-259, Department of Computer Science, University of California, Los Angeles, 1998.

[Shapley and Shubik, 1954] Lloyd S. Shapley and Martin Shubik. A method for evaluating the distribution of power in a committee system. *The American Political Science Review*, 48(3):787–792, 1954.

[Shapley, 2016] Lloyd S. Shapley. *17. A value for n-person games*. Princeton University Press, 2016.

[Shoham and Leyton-Brown, 2008] Yoav Shoham and Kevin Leyton-Brown. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

[Talbert, 2025] Matthew Talbert. Moral Responsibility. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2025 edition, 2025.

[Tigard, 2021] Daniel W Tigard. Artificial moral responsibility: How we can and cannot hold machines responsible. *Cambridge Quarterly of Healthcare Ethics*, 30(3):435–447, 2021.

[Tognazzini and Coates, 2025] Neal Tognazzini and D. Justin Coates. Blame. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2025 edition, 2025.

[Triantafyllou and Radanovic, 2023] Stelios Triantafyllou and Goran Radanovic. Towards computationally efficient responsibility attribution in decentralized partially observable mdps. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 131–139, 2023.

[Triantafyllou *et al.*, 2021] Stelios Triantafyllou, Adish Singla, and Goran Radanovic. On blame attribution for accountable multi-agent sequential decision making. *Advances in Neural Information Processing Systems*, 34, 2021.

[Triantafyllou *et al.*, 2022] Stelios Triantafyllou, Adish Singla, and Goran Radanovic. Actual causality and responsibility attribution in decentralized partially observable markov decision processes. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 739–752, 2022.

[Tsirtsis *et al.*, 2024] Stratis Tsirtsis, Manuel Gomez Rodriguez, and Tobias Gerstenberg. Towards a computational model of responsibility judgments in sequential human-ai collaboration. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.

[Van de Poel, 2015] Ibo Van de Poel. Moral responsibility. In *Moral responsibility and the problem of many hands*, pages 12–49. Routledge, 2015.