CESifo WORKING PAPERS

9898
2022

August 2022

A Theory of Causal Responsibility Attribution

Florian Engl

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 9898*

A Theory of Causal Responsibility Attribution

Abstract

People often act out of a desire to be responsible for good and not for bad events. Similarly, people frequently reward and punish other people if they perceive them to be responsible for the implementation of events that they like or dislike. When the implementation of an event depends on the interaction of multiple persons and, potentially, moves of nature, the determinants of such responsibility perceptions are not well understood. In this paper, I propose a notion of causal responsibility which attempts to objectively capture the causal importance of a person's action for the implementation of an event in such situations. A laboratory experiment shows that the notion successfully predicts people's responsibility perceptions. Furthermore, I incorporate the notion in a framework of responsibility preferences and study its implications for worker motivation and the design of voting rules. Finally, I show that the notion can explain experimentally elicited behavior and punishment and reward patterns in multi-agent situations that are not well-explained by existing social preference theories.

JEL-Codes: C720, D030, D630, D700.

Keywords: responsibility, causal reasoning, social preferences.

Florian Engl
University of Regensburg / Germany
florian.engl@ur.de

August 2022
I want to thank Björn Bartling, Ernesto Dal Bo, Pedro Dal Bo, Lea Cassar, Guillaume Fréchette, Bernd Irlenbusch, Matthew Jackson, Dorothea Kuebler, Igor Letina, Nick Netzer, Axel Ockenfels, Arno Riedl, María Sáez-Martí, Alexander Sebald, Andrew Schotter, Dirk Sliwka, Joel Sobel, Ran Spiegler, and Roberto A. Weber and numerous seminar and conference participants for very helpful guidance and comments. I want to thank the Russell Sage Foundation (award #98-15-06) and the Dr.-Jürgen-Meyer-Foundation for financial assistance. The laboratory experiment has been conducted according to the ethical guidelines of the Cologne Laboratory for Experimental Research (CLER) of the University of Cologne.

"The causes of pain and pleasure, whatever they are, or however they operate, seem to be the objects, which (...) immediately excite those two passions of gratitude and resentment."
— Adam Smith, *The Theory of Moral Sentiments*, 1759, p. 84

1 Introduction

Perceptions of responsibility are a pervasive component of everyday life. Many people's actions are influenced by their desire to feel responsible for good and to avoid feeling responsible for bad events. Similarly, in social interactions, people often praise or blame other people if they perceive them to be responsible for good or bad events. Through these channels, responsibility perceptions can play an important role in many economic environments. For example, they can influence a manager's decision to fire a worker who he perceives to be responsible for the failure of a project. Governments negotiate over who is responsible for climate change and how those not responsible should be compensated. Voters attribute responsibility to parties or individual politicians for the implementation of reforms or the state of the economy and vote to reward or punish those actors. In markets, the question arises whether firms or their customers are responsible for negative externalities of the production process. Ultimately, institutions like hierarchies and voting rules can influence responsibility perceptions, which should be taken into account when designing them.[1]

Despite this importance, an established and widely applicable theoretic notion of responsibility does not exist within economics. The goal of this paper is, therefore, to provide a notion of responsibility that can serve as a benchmark to study such situations. Specifically, I formalize, provide evidence for, and explore the strategic implications of a notion of responsibility which attempts to capture objectively the causal importance of an agent's action for the implementation of an event: *causal responsibility*.[2]

While the evaluation of causal responsibility is straightforward when only one actor is involved, it soon becomes complicated when multiple parties and, potentially, moves of nature interact and, thus, the question arises who is more or less responsible for an event's implementation. An early philosophical example for this complexity is the desert-traveler paradox: A traveler who passes a desert has two enemies. Both want to kill him. The first enemy poisons the traveler's water. The second enemy empties the traveler's water bottle. The traveler consequently dies of thirst. Who is the cause of his death? Who is responsible? Are both enemies as responsible as a single actor would have been? Or are they both not responsible at all, because

[1] A growing literature in economics uses experiments to study the implications of responsibility perceptions, e.g., for workers' effort provision (Charness, 2000, 2004), delegation (Hamman et al., 2010; Bartling and Fischbacher, 2012; Gawn and Innes, 2021), voting decisions (Bartling et al., 2015a; Duch et al., 2014; Anselm et al., 2022), the choice of unethical outcomes (Behnk et al., 2019; Falk et al., 2020), human-machine interactions (Kirchkamp and Strobel, 2019), behavior within markets (Bartling et al., 2015b; Kirchler et al., 2016), dictator game giving (Cryder and Loewenstein, 2012), whistleblowing (Choo et al., 2019), endogenous group formation (Brütt et al., 2020), and redistributive preferences (Cappelen et al., 2010, 2020). In the theoretical work of Prendergast (1995) and Sliwka (2006), responsibility is allocated ex ante by a principal to a worker for the implementation of a task, Manove (1997) models "responsible jobs" as those in which a worker's effort can influence the output, and Tungodden (2005) and Cappelen and Tungodden (2006) model personal responsibility as a determinant of redistribution.

[2] Of course, different forms of responsibility exist. For example, Hart (1968) categorizes four different notions of responsibility that play a role in legal contexts: role responsibility, causal responsibility, liability responsibility, and capacity responsibility. I focus on causal responsibility because it provides a fundamental, objective basis for responsibility evaluations.

the traveller would have died anyway from the other enemy's action?

In my approach, an agent's *overall causal responsibility* for the implementation of an event is assessed using counterfactual and probabilistic reasoning, which is reflected in a convex combination of an ex post and an ex ante causal responsibility part. *Ex post causal responsibility* is based on a simple principle: it decreases the further away an agent's strategy is from being pivotal for the implementation of an event. An agent is pivotal for an event, if he had a different strategy at his disposal that would have, ceteris paribus, prevented the event's implementation. "Distance" from pivotality is measured by the expected minimum number of changes that would need to be made to the other agents' strategies in order to make the agent under consideration pivotal. Finally, if an agent's strategy could never be pivotal for an event, he bears no ex post causal responsibility for it.[3] Under these assumptions the two enemies of the traveller each have partial causal responsibility for the traveller's death as each of them could have been pivotal if the other enemy had acted differently. This coincides with how many people would intuitively judge this situation. An important consequence of the theory is that it predicts a difference in responsibility attribution between agents that act as substitutes and those that act as complements for an event. While ex post causal responsibility is diffused among substitutes, such as the two enemies of the traveler, it is not diffused among complements, such as a group of coauthors or a management team, for which the overall success of a project crucially depends on each members' contribution.[4]

Ex ante causal responsibility captures that there can exist uncertainty—either from moves of nature or the agents' strategies—about the degree of ex post causal responsibility that follows from a given strategy. For example, consider two persons who both shoot a gun but have very different chances of hurting an innocent bystander. Both are fully ex post causal responsibility if the bystander is hurt but, intuitively, the shooter with the higher ex ante probability bears higher responsibility. Ex ante causal responsibility is defined as the expected level of ex post causal responsibility evaluated at the point at which an agent makes his first decision. It can therefore capture this intuition. Through ex ante causal responsibility, it is also possible to assign responsibility for events that persons could have been pivotal for, even if these events are never actually implemented. For example, if a person drives recklessly or under the influence of drugs but doesn't hurt anybody, ex ante causal responsibility captures that there was a high chance of hurting somebody and that the driver should be held to some extent responsible for that hypothetical event.

For itself, the notion of causal responsibility is a useful theoretic tool for assessing responsibility in multi-agent situation. To test whether the theory is also successful in making correct comparative-statics prediction about the formation of actual responsibility perceptions, I conducted an incentivized laboratory experiment. In the experiment, participants are presented with four different hypothetical scenarios in which two persons interact to implement an outcome, and are asked to rate the responsibility of one of the persons for the implementation of the outcome. The four scenarios vary the pillar of causal responsibility that is tested (ex

[3]Several recent papers highlight the importance of pivotality for decision-making (Bénabou et al., 2018; Rothenhäusler et al., 2018; Falk et al., 2020). However, they don't consider distance from pivotality, which can make a difference in many applications.

[4]This feature differentiates my notion from naive diffusion of responsibility, which simply divides responsibility among the involved actors independent of their causal impact (cf. Darley and Latané, 1968).

ante and ex post) and the strategic setting (simultaneous or sequential move). Each scenario is presented in two variations that only differ in one respect such that the notion of causal responsibility makes a clear comparative-statics prediction regarding the difference in responsibility perceptions between the two variants. Each scenario yields a significant result confirming the prediction of the theory and, thus, provides support for the assumptions behind the notion of causal responsibility.

An important feature of the formal notion of causal responsibility is that it can be incorporated in different preference frameworks. For example, people can care about their intrinsic feeling of causal responsibility for an event, or about attributing causal responsibility to other people.[5] In this paper, I first incorporate the notion in a framework of preferences for *internal responsibility attribution* which implies that, in addition to preferences over monetary payoffs, agents have a preference to (not) be causally responsible for the implementation of events that they judge as good (bad). I state the conditions for a sequential responsibility equilibrium in which every actor chooses a strategy that, at each history, maximizes his utility and beliefs are correct.

In an application, I show how preferences for internal causal responsibility attribution can influence workers' effort provision. Specifically, workers can work or shirk on day one of a two-day project. The project is only completed early if all team members work on day one, otherwise it is completed late. Under standard preferences of pure self-interest, all workers would shirk and delay the completion of the project to day two. However, if workers care enough about their causal responsibility for the delayed completion, there exists an equilibrium in which the project is completed early.[6]

I show that, with homogeneous workers, increasing the team size diffuses causal responsibility for the delay and thus increases the range of parameters for which a "shirking"-equilibrium is sustained. Furthermore, holding the team size constant, I show that, with heterogeneous workers, the distribution of the workers' concern for causal responsibility influences the implemented equilibrium. For example, I demonstrate that one worker who has no concern for causal responsibility and who will therefore always shirk, can start a cascade of responsibility diffusion at the end of which even workers with a very high concern for causal responsibility shirk. On the other hand, I show that one worker with high concern for causal responsibility can also inspire a whole team to work and complete the project early. Finally, I show that outcome- and intention-based social preference models either make predictions that are opposed to those of the causal responsibility-based model, or are indistinguishable from the standard model.

To test whether internal causal responsibility attribution actually affects behavior as predicted, I consult evidence from existing experiments. Specifically, I use the study by Falk et al. (2020) on diffused pivotality. In their study, a negative externality is implemented if at least one out of eight group members chooses an option that increases his or her own monetary payoff. Thus, a group members' causal responsibility for the negative externality decreases the more

[5] In addition, the notion could be applied as a moderator of other emotions like reciprocity (Rabin, 1993), frustration and anger (Battigalli et al., 2016), or guilt (Battigalli and Dufwenberg, 2007).

[6] Internal causal responsibility attribution thus acts as a non-monetary incentive. Other examples of non-monetary are an individual's desire to perform a task for its own sake (Bénabou and Tirole, 2003; Cassar and Meier, 2018), a social norm (Sliwka, 2007), pride and self-esteem (Ellingsen and Johannesson, 2008), or reciprocity concerns (Von Siemens, 2013).

other group members also choose the payoff-maximizing option. Hence, internal causal responsibility attribution predicts that the likelihood with which the payoff-maximizing option is chosen by a specific group member increases with the number of other group members that that specific group member believes will also take that option. Since the study also elicits each group members' beliefs about how many other group members take the payoff-maximizing action, we can directly test the predictions of the theory. And indeed, in accordance with the theory, the probability of choosing the payoff-maximizing action significantly increases, the more other group members are believed to also take it.

I also incorporate the notion in a framework of preferences for *external causal responsibility attribution* which implies that, in addition to preferences over monetary payoffs, an agent can have a taste to reward or punish other agents for the implementation of what he or she judges as good or bad events, but only to the extent that those agents are causally responsible for the event. I analyze the strategic implications of such preferences in simple two-stage games. In the *collective-action stage*, several agents with standard preferences and, potentially, nature collectively implement an event. Thereafter, in the *responsibility-attribution stage*, a separate agent with responsibility-based preferences evaluates the possible events and the agents' causal responsibility for them. He or she then has the opportunity to punish or reward the agents through an allocation decision. This simple setup allows us to capture many scenarios that are typically associated with the allocation of responsibility. In a sequential responsibility equilibrium, agents anticipate punishment and reward and will therefore seek to evade responsibility for events they think are judged as bad and seek responsibility for events they think are judged as good.

In an application, I analyze how external causal responsibility attribution matters for the design of voting rules. In particular, I study a situation in which the preferences of the members of a committee and the stakeholders of the committee are misaligned. For example, while a company's board of directors is supposed to represent the interests of the shareholders, it is not unlikely that conflicts-of-interest arise, e.g., due to the common dual role of board chairman and chief executive. In the *collective-action* stage, members of the committee vote for or against the implementation of an event that is beneficial for the committee's stakeholders. However, committee members get a private payoff if the event is not implemented.[7] In the *responsibility-attribution stage*, the stakeholders, at some cost, reward or punish the committee members for their voting behavior. The sequential Nash equilibrium predictions of standard preferences are clear: Since punishment and reward is costly, the stakeholders will never engage in it and, thus, in the first stage, the committee will vote against the beneficial event, independent of the specific voting rule. However, if the stakeholders have responsibility preferences, equilibria in which the event is implemented are possible. I show how the existence of such an equilibrium in majority voting depends on the number of committee members and demonstrate that a consensus rule is better for ensuring that implementing the event is the unique responsibility equilibrium.

Of course, punishment or reward can be motivated by many different reasons. Whether responsibility perceptions are an important reasons, is, in the end, an empirical question. Therefore, I show in a final part that the comparative-statics predictions of the theory are consistent

[7] In the case of the board of directors, it could be a vote on the compensation package for the executive team.

with existing experimental evidence of punishment behavior. For example, causal responsibility attribution can explain experimental evidence on the blame-reduction potential of delegation, the outcome bias, strategic voting, or redistribution decisions. Throughout the analyses, I compare the predictions of causal responsibility attribution to those of standard preferences and other social preference theories. Causal responsibility explains observed punishment patterns more successfully than existing theories and remains, in regression analyses, a highly significant predictor for punishment even after controlling for several other potential motives.

The paper that is most related is the experimental study of Bartling and Fischbacher (2012), which also develops a theoretic notion of responsibility in a multi-agent context. Their notion attributes responsibility for an event to an agent in proportion to the increase in the event's likelihood that is due to that agent's action, compared to ex ante beliefs about the likelihood. The two approaches differ considerably. In their approach, the responsibilities of all agents always sum up to one. Therefore, responsibility is diffused among the involved agents, independent of whether agents are pivotal or not. Furthermore, their measure ignores moves of nature, while mine specifically recognizes their role in the formation of causal responsibility perceptions. Finally, their notion is not incorporated in a preference framework or a game-theoretic equilibrium concept. However, their notion crucially depends on the ex ante beliefs of the agent who evaluates responsibility. Since beliefs are not stipulated by an equilibrium concept, they are assumed to concord with "average play", which can lead to counterintuitive predictions. For example, if average play is that an action is taken with certainty, as would be the case in a pure-strategy equilibrium, no agent who takes that action is attributed any responsibility for the event that follows, even if his action is pivotal.

The remainder of the paper is organized as follows. In Section 2, I review evidence from a variety of domains supporting the existence of a link between causality, responsibility, and the attribution of blame and praise. Section 3 introduces the formal notion and Section 4 provides experimental evidence. Section 5 incorporates the notion in a preference framework for internal responsibility attribution, and Section 6 does the same for preferences for external responsibility attribution. Finally, Section 7 concludes and lays out avenues for future research.

2 Related literature on causal responsibility

In this section, I review evidence from a variety of domains supporting the existence of a link between causality, responsibility, and the attribution of blame and praise. A first part addresses the difficulty of stipulating what counts as a cause and reviews evidence suggesting that there is a natural tendency in people to think in causal frameworks and to use counterfactual as well as probabilistic reasoning to do so. The second part connects causal reasoning to perceptions of responsibility and the attribution of blame and praise.

2.1 Causation, counterfactuals, and probabilities

Causal reasoning is a pervasive component of people's everyday mental life (Summerville and Roese, 2008) and the capacity to perceive causation already exists in infants (Leslie and Keeble, 1987; Saxe and Carey, 2006). Reviewing evidence of people's tendency to formulate causal

explanations for surprising events, Tversky and Kahneman (1982, p. 128) conclude that they "were impressed by the fluency which our respondents displayed in developing causal accounts" and, consequently, Kahneman and Frederick (2002) categorize causal reasoning as belonging to System-1, i.e., to be an intuitive, automatic, and effortless mental process.[8]

Counterfactual reasoning and ex post causal responsibility. However, what constitutes a cause has been the object of intense debate in the theoretical causality literature. Discussions of counterfactual reasoning as a tool to assess causality date back to, at least, David Hume, the 18th-century philosopher, who defined "a cause to be an object, followed by another, (...) where, if the first object had not existed, the second never had existed." (Hume, 1777, Section VII, Part II). Such standard counterfactual reasoning (c.f. Lewis, 1973, 1986) can successfully assess causation in simple situations when an action is pivotal for an event, but fails when events are causally *overdetermined* as in the desert-traveler paradox. Several approaches have been proposed to attribute causation even in such cases where no single action is pivotal. I'm following the widely-accepted *structural* approaches, which allow a non-pivotal action to be a cause of an event, if it could have been pivotal for the event under some counterfactual changes to the environment (cf. Pearl, 2000; Woodward, 2003; Halpern and Pearl, 2005).[9]

Psychological research has demonstrated that counterfactual reasoning is not only a useful theoretic tool, but also important for how people actually perceive causality. For example, Kahneman and Tversky (1982) proposed that people use mental simulations of counterfactuals when making causal inferences, which inspired much of the subsequent psychological research on causal reasoning (e.g., Wells and Gavanski, 1989; Gerstenberg et al., 2014; Roese, 1997, for an overview). In a clever study Gerstenberg et al. (2017) use eye-tracking to provide direct evidence that people's causal judgments are influenced by counterfactuals. They show that eye movements track potential counterfactual movements of billiard balls on a computer screen before causal judgments about those billiard balls are made.

To describe the relationship between counterfactual causal reasoning and ex post causal responsibility formally, I build on a notion of responsibility that was pioneered by Chockler and Halpern (2004) in the artificial intelligence literature. In Chockler and Halpern's notion, the responsibility of A for a realized event B inversely depends on the minimum number of changes that have to be made to the specific context in order to make A pivotal for event B. The predictive power of this relationship for responsibility perceptions has been tested and confirmed in a series of studies by cognitive psychologists (e.g., Gerstenberg and Lagnado, 2010; Zultan et al., 2012; Lagnado et al., 2013).

Probabilistic reasoning and ex ante causal responsibility. In addition to counterfactual reasoning, Kahneman and Varey (1990) developed the idea that people's perception of causality is also inherently related to judgments of probabilities about the counterfactuals. Empirical evidence supports probabilistic reasoning as a second pillar for causal reasoning (see, e.g.,

[8]This does not mean, however, that people always draw the correct causal conclusions when engaging in causal reasoning. Indeed, much of the early literature focussed on potential biases in the assessment of causality, e.g., Michotte's classic studies on people's tendency to see causes where there are none (Michotte, 1963).

[9]Other approaches attack the problem by defining an action as a cause if it is a "necessary element of a set of conditions jointly sufficient for the result" (NESS test, c.f., Hart and Honoré, 1959; Wright, 1985), or if it is an "insufficient but necessary part of a condition which is itself unnecessary but sufficient" (INUS condition, c.f., Mackie, 1974). For an extensive recent summary of different approaches to causality, see Beebee et al. (2012).

Mandel and Lehman, 1996; McGill and Tenbrunsel, 2000). Such ex ante causality ratings are also discussed in Spellman (1997); Gerstenberg and Lagnado (2010, 2012); Zultan et al. (2012); Lagnado et al. (2013); Gerstenberg et al. (2018). My approach of causal responsibility, which is formally introduced in the next section, combines counterfactual and probabilistic reasoning as determinants of responsibility perceptions and thus closely follows the psychological evidence.

2.2 Causal responsibility as a determinant of blame and praise

Finally, perceptions of causality have been found to be an important determinant of blame and praise. For example, Adam Smith believed that people exhibit an innate preference to show gratitude and resentment for the causes of one's pain and pleasure. For him, the reaction to look first and immediately for the cause of one's fortune was so innate that he believed that animals and even inanimate objects, like stones, could become a person's target for gratitude and resentment if they *caused* that person pain or pleasure—something that clearly cannot be explained by standard preferences or intentions (Smith, 1759, p. 85). The relationship between causal responsibility and punishment is also present in the legal system. While legal economists often emphasize the incentive and social welfare effect of legal liability in law and thereby neglect the question of actual causation (Coase, 1960; Landes and Posner, 1983), many legal scholars argue that legal liability should reflect a defendant's underlying moral responsibility and that moral responsibility is based on causation. This holds especially in criminal and tort law, where the retributive and corrective justice effect of punishment is emphasized. Consequently, starting with Hart and Honoré (1959) and Hart (1968), many legal scholars have studied the relationship between causation, responsibility, and legal liability (Epstein, 1973; Wright, 1985, 1988; Moore, 2009; Green, 2015; Steel, 2015).

A large literature in psychology supports the assumption that perceptions of causality and responsibility are an important driver for the attribution of blame and praise.[10] For example, the connection has been discussed widely within the framework of attribution theory (Heider, 1958; Kelley, 1967; Shaver, 1985; Weiner, 1995). Other theories propose causation as the first necessary condition on a path towards the decision to blame and punish (Darley and Shultz, 1990; Malle et al., 2014). For example, Darley and Shultz (1990) state that "judgments of moral responsibility presuppose those of causation. If the protagonist is judged not to have caused the harm, then there is no need to consider whether he is morally responsible for it. Similarly, judgments of blame presuppose those of moral responsibility. And finally, decisions about punishment presuppose judgments of blame." Using mostly vignette studies, psychologists have also empirically shown a positive correlation between judgments of causality and responsibility (Wells and Gavanski, 1989), and that perception of causation influence perceptions of responsibility which, in turn, influence punishment (Shultz et al., 1981). Cushman (2008) provides evidence that shows that, while the perceived moral wrongness of an action is determined by the actor's intention, attributed blame and punishment for that action is more sensitive to perceptions of causal responsibility.

[10] Alicke et al. (2015) state in a historical overviews of psychological models of blame assignment and the role of causality therein that "[c]ausation is the link that cements (...) prior mental states to the outcomes that ensue and, along with intention, is the primary criterion for evaluating social behavior." Guglielmo (2015) provides another overview.

3 A notion of causal responsibility

This section formalizes a notion of causal responsibility based on the aforementioned principles. Theoretically, I build on the notation of Battigalli and Dufwenberg (2009) and Battigalli et al. (2016)'s framework of dynamic psychological games. Using psychological game theory is necessary as beliefs directly enter an agent's utility function through their role in the evaluation of causal responsibility. The general framework in which responsibility is evaluated is a finite multi-stage environment with potential moves of nature. In each stage, each agent i) chooses exactly one discrete action, ii) knows all preceding choices, and iii) has no knowledge of the other agents' choices in that stage. Hence, all instances of imperfect information arise from simultaneous moves. It is also possible that some agents are inactive in a given stage (then their action set is a singleton). In perfect information games, in each stage, exactly one agent is active and all others are inactive. Thus, this framework encapsulates simultaneous move games, sequential move games, and repeated games as special cases.

Notationally, I define an *extensive form* as a tuple $\langle I, H, Z \rangle$ of agents and histories. Agents are indexed by $i \in I = \{1, ..., n\}$. When moves of nature are possible, an augmented set of agents $I_c = I \cup \{c\}$, where c denotes nature, is considered. H and Z denote the finite sets of nonterminal and terminal histories, respectively. Histories are sequences of choices made by agents on the path of the game. For example, $h = (a^1, ..., a^t)$ is a history of length $t > 1$ where $a^t = (a_i^t)_{i \in I_c}$ denotes the chosen action profile at stage t. The empty history $h = \varnothing$ is the root of the game. Nonterminal histories can be interpreted as decision nodes and terminal histories as endnodes of the game.

After each history $h \in H$, each agent i chooses an action a_i from his finite set of feasible actions, $A_i(h)$, which can depend on history h. S_i and Σ_i denote agent i's sets of pure and behavior strategies, respectively. A pure strategy $s_i \in S_i$ prescribes, for each history $h \in H$, the choice of an action $a_i \in A_i(h)$. A behavior strategy $\sigma_i \in \Sigma_i$ assigns, for each history $h \in H$, a probability distribution over the set of feasible actions, $A_i(h)$. If nature moves at history $h \in H$, her move is prescribed by a commonly known probability distribution, $\sigma_c(a_c|h) \in \Delta(A_c(h))$, over her potential actions at history h.

The set $S(h) \subseteq S$ comprises all pure strategy profiles that allow history h. Let $\sigma_i(h)$ be agent i's *updated* strategy that prescribes the same probabilities to actions as σ_i, except for the actions that define history h, which are taken with probability 1. Thus, the *updated* strategy profile $\sigma(h) = (\sigma_i(h))_{i \in I_c}$ comprises what has already happened and what is still uncertain at history h. $\Sigma(h)$ denotes the corresponding set of updated behavior strategy profiles.

Every terminal history is uniquely determined by a pure strategy profile, $z : S \to Z$. Terminal histories determine events $x \in X$ according to the function $f : Z \to X$. It is thus possible that the same event is implemented by several terminal histories. This can happen, for example, when the implementation of an event is determined by majority voting.

3.1 Distance from pivotality

In the following, I present five simple steps that guide the evaluation of causal responsibility. The key concept is that an agent i's causal responsibility for an event x decreases with the

expected "distance" that his actions are away from being pivotal for the implementation of event x. To formally derive an agent i's distance from pivotality for event x, I define, as a first step, a *pivotality set* $\tilde{S}_{i,x}(s_i, \mathbf{s_{-i}})$). This set comprises all those strategy profiles for which agent i's strategy s_i would be pivotal for event x and is formally defined as follows:

Definition 1. *The* pivotality set *of agent i for any event $x \in X$ and any pure strategy profile $(s_i, \mathbf{s_{-i}}) \in S$ is defined as*

$$\begin{aligned}\tilde{S}_{i,x}(s_i, \mathbf{s_{-i}}) =& \{(s_i, \tilde{\mathbf{s}}_{\mathbf{-i}}) \in S(h_i(s_i, \mathbf{s_{-i}})) \quad \textit{such that} \\ & (1)\ f(s_i, \mathbf{s}_{-i}) = f(s_i, \tilde{\mathbf{s}}_{-i}) = x \\ & (2)\ \exists\, s_i' \in S_i\ s.t.\ f(s_i', \tilde{\mathbf{s}}_{-i}) \neq x\}.\end{aligned} \tag{1}$$

The pivotality set $\tilde{S}_{i,x}(s_i, \mathbf{s_{-i}})$ thus comprises all pure strategy profiles $(s_i, \tilde{\mathbf{s}}_{\mathbf{-i}}) \in S(h_i(s_i, \mathbf{s_{-i}}))$ for which two conditions hold: First, a strategy profile $(s_i, \tilde{\mathbf{s}}_{\mathbf{-i}})$ has to implement the considered event x, which is also implemented by the actual strategy profile $(s_i, \mathbf{s_{-i}})$. Second, if the other agents choose the strategy profile $\tilde{\mathbf{s}}_{-i}$, agent i has to have an alternative strategy s_i' that, if taken, would change the event. In other words, if the others choose $\tilde{\mathbf{s}}_{\mathbf{-i}}$, then agent i's actual strategy s_i is pivotal for event x. To avoid that agent i is held responsible for events that realize in parts of the game tree that are never reached given the actual strategy profile $(s_i, \mathbf{s_{-i}})$, I introduce $h_i(s_i, \mathbf{s_{-i}}) \in H$ as the history at which agent i is active for the first time given $(s_i, \mathbf{s_{-i}})$ and restrict the set of potential strategy profiles, $S(h_i(s_i, \mathbf{s_{-i}}))$, to those that lead to agent i's first action.

Example 1 *(Voting). Step 1:* Consider a hiring committee with five members, $I = \{1, ..., 5\}$, who vote by majority rule on whether to hire a specific applicant or not ($A_i = \{y, n\}$, $X = \{hire, not\ hire\}$). Assume that four members vote for and one member against hiring the applicant ($\mathbf{s} = (y, y, y, y, n)$). Thus, she will be employed ($f(y, y, y, y, n) = hire$).

Step 1: The *pivotality set* of committee member 1 given the actual event *hire* and strategy profile (y, y, y, y, n) is $\tilde{S}_{1,hire}(y, y, y, y, n) = \{(y, n, y, y, n), (y, y, n, y, n), (y, y, y, n, n),\ (y, y, n, n, y),\ (y, n, y, n, y),\ (y, n, n, y, y)\}$. ▲

After having identified all those strategy profiles for which agent i would be pivotal for event x, the second step is to determine how far he is away from being pivotal given the considered strategy profile $(s_i, \mathbf{s_{-i}})$. To this end, I define a function $c(\mathbf{s}_{-i}, \tilde{\mathbf{s}}_{-i})$ which simply measures the differences between strategy profiles $\mathbf{s}_{-i}$ and $\tilde{\mathbf{s}}_{-i}$.

Definition 2. *The number of differences between strategy profiles $(s_i, \mathbf{s_{-i}}) \in S$ and $(s_i, \tilde{\mathbf{s}}_{\mathbf{-i}}) \in \tilde{S}_{i,x}(s_i, \mathbf{s_{-i}})$ is defined as*

$$c(\mathbf{s}_{-i}, \tilde{\mathbf{s}}_{-i}) = \sum_{j \in I_c \setminus \{i\}} \mathbb{1}(s_j \neq \tilde{s}_j) \tag{2}$$

The function $c(\mathbf{s}_{-i}, \tilde{\mathbf{s}}_{-i})$ counts the number of strategies that differ between the actual profile $\mathbf{s}_{-i}$ and the profile $\tilde{\mathbf{s}}_{-i}$ that sets agent i's strategy pivotal. In other words, the function counts *how many* of the other agents would need to change their strategy in order to make agent i's strategy pivotal for event x.

Example 1 *(Voting - cont'd). Step 2:* The number of differences between the chosen voting strategy profile of the *other* agents, (y, y, y, n), and the profiles in the *pivotality set* of committee member 1 differ. For example, $c((y, y, y, n), (n, y, y, n)) = 1$ and $c((y, y, y, n), (n, n, y, n)) = 2$. ▲

Finally, distance from pivotality is measured by the minimum number of changes that the other agents would need to make to their strategies such that agent i becomes pivotal for event x. Formally, distance from pivotality is defined as follows:

Definition 3. *The distance of agent i from being pivotal for event x for any pure strategy profile* $(s_i, \mathbf{s_{-i}}) \in S$ *is defined as*

$$d_{i,x}(s_i, \mathbf{s_{-i}}) = \begin{cases} \min_{\tilde{\mathbf{s}} \in \tilde{S}_{i,x}(s_i, \mathbf{s_{-i}})} c(\mathbf{s}_{-i}, \tilde{\mathbf{s}}_{-i}) & \text{if} \quad \tilde{S}_{i,x}(s_i, \mathbf{s_{-i}}) \neq \emptyset \\ \infty & \text{if} \quad \tilde{S}_{i,x}(s_i, \mathbf{s_{-i}}) = \emptyset. \end{cases} \tag{3}$$

Thus, the minimum possible distance between $\mathbf{s}_{-i}$ and $\tilde{\mathbf{s}}_{-i}$ is 0, which means that agent i is already pivotal given the actual strategy profile. The maximum possible distance is set to infinity, which is realized if there exists no alternative strategy profile which would set agent i pivotal for event x.

Example 1 *(Voting - cont'd). Step 3:* The minimum distance of member 1 from being pivotal for the hire is 1, $d_{1,hire}(y, y, y, y, n) = 1$. ▲

Note that the distance function could be extended to take other considerations into account, e.g., how costly the agents' changes are. Investigating such additional components could influence overall responsibility perceptions and is therefore certainly worthwhile for future research. However, it lies beyond the scope of this paper, namely to develop a concept of *causal* responsibility. Furthermore, the notion is primarily built to evaluate responsibility when actions with nominal interpretation are taken. Actions with cardinal interpretation, i.e., action in which the question arise *how much* an agent's strategy contributed to an outcome, require an adjustment of the distance function.

3.2 Ex post causal responsibility

The function $r_{i,x}^{EP}$ denotes agent i's degree of ex post causal responsibility for event x. In order to capture the intuitions in the introductions, the minimum requirement is that $r_{i,x}^{EP}$ is decreasing and convex in distance from pivotality (i.e., $\frac{\partial r_{i,x}^{EP}}{\partial d_{i,x}} < 0$ and $\frac{\partial^2 r_{i,x}^{EP}}{\partial d_{i,x}^2} > 0$). To increase tractability and ease of application, I use a specific functional form to define ex post causal responsibility, namely the inverse of the distance from pivotality (cf. Chockler and Halpern, 2004).

Definition 4. *An agent i's degree of ex post causal responsibility for any event* $x \in X$ *and any updated behavioral strategy profile* $\sigma(z) \in \Sigma(z)$ *is defined as*

$$r_{i,x}^{EP}(\sigma(z)) = \mathbb{E}[\frac{1}{1 + d_{i,x}(s_i, \mathbf{s_{-i}})} | \sigma(z)] \tag{4}$$

where

$$\mathbb{E}[\frac{1}{1 + d_{i,x}(s_i, \mathbf{s_{-i}})} | \sigma(z)] = \sum_{\mathbf{s} \in S(z)} Pr(\mathbf{s} | \sigma(z)) \; \frac{1}{1 + d_{i,x}(s_i, \mathbf{s_{-i}})}.$$

So far, all components of the responsibility function were defined in terms of pure strategy profiles. However, if agents play behavior strategies, it has to be taken into account that, aside from the path of the game that was actually played, uncertainty can remain about which actions would have been played off that path (see Example 4).[11] Because distance from pivotality can depend on the uncertain actions that are off the realized history, this uncertainty can influence ex post causal responsibility values. Therefore, the ex post causal responsibility measure is an expected value where $Pr(\mathbf{s}|\sigma(z))$ is the probability that the fully updated behavior strategy profile for terminal history z, $\sigma(z)$, assigns to pure strategy profile $\mathbf{s}$.

Ex post causal responsibility values are easy to calculate and have an intuitive interpretation. The function is bounded as its value can only lie in the interval from 0 to 1. The interval can be interpreted as no, partial, and full ex post causal responsibility. Furthermore, the function monotonously decreases in the distance from pivotality. Finally, ex post causal responsibility values are robust to changes in the game form. For example, as distance is defined by changes to strategies instead of actions, adding an action that increases the number of actions which need to be changed to reach pivotality but is otherwise irrelevant, does not alter the degree of responsibility. The same holds for the addition of irrelevant actors. As long as an added player does not alter the pivotality set, responsibility levels are unchanged. The following examples demonstrate the versatility of ex post causal responsibility:

Example 1 *(Voting - cont'd). Step 3:* The ex post causal responsibility of member 1 for the hire is $r^{EP}_{1,hire}(y,y,y,y,n) = \frac{1}{1+1} = \frac{1}{2}$. If only three committee members would have voted yes, all of them would have full ex post causal responsibility would have been $r^{EP}_{1,hire}(y,y,y,n,n) = 1$.▲

Example 2 *(Sequential moves).* In this example, agents move sequentially. Alex can either insult Bea or not. If he doesn't insult her, there is no fight and the game ends. If he insults her, she can react by retaliating or by ignoring him. If she ignores him, there is no fight but if she retaliates the two start a fight. Figure 1 shows the extensive form of this game:

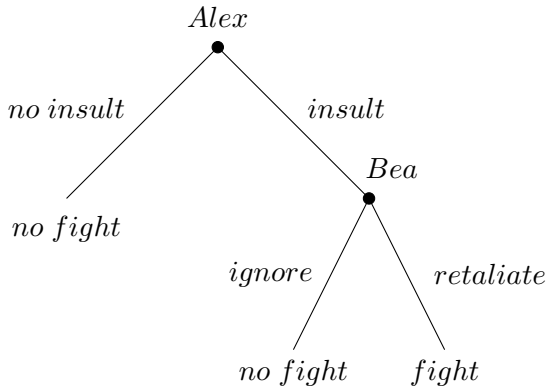


Figure 1: Ex post causal responsibility with sequential moves.

What are Alex's and Bea's ex post causal responsibility measures for the different outcomes. First, consider the case in which Alex provokes, Bea retaliates and a fight starts. Both are fully ex post causally responsible for the fight because they both could have prevented it. However,

[11] That is because with behavior strategies, as opposed to mixed strategies, the chosen actions are not determined at the start of the game, but at each decision node.

if Bea remains calm and ignores Alex' insult, only she is fully ex post causally responsible for avoiding the fight. After provoking, Alex cannot have positive ex post causal responsibility for *no fight* as he cannot be pivotal for it.

When Alex doesn't insult Bea, his ex post causal responsibility for *no fight* stochastically depends on what Bea would have done after a provocation. If she retaliates with probability p, then his ex post causxal responsibility is $r^{EP}_{Alex,no\ fight}(no\ insult, p) = p * 1 + (1-p) * \frac{1}{2}$. That is, Alex will only get full ex post causal responsibility for not starting a fight, if Bea would have retaliated with certainty. ▲

Example 3 *(Multiple moves).* Causal responsibility can also be evaluated when agents move multiple times along the same path of the game. Consider the following adaptation of the centipede game. The fair event is only implemented if Chris and Dan always choose right (R or r, respectively). If one of them chooses down (D or d, respectively), the game ends and an unfair event materializes.

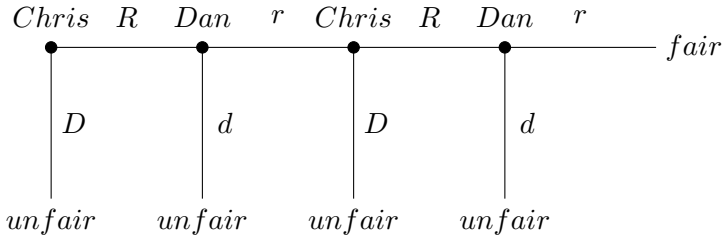


Figure 2: Ex post causal responsibility with multiple moves.

Imagine Chris and Dan both choose right twice. In this case, they are both pivotal and, hence, fully ex post causally responsible for the fair event ($r^{EP}_{Chris,fair}(RR, rr) = 1$). When Chris always chooses down and Dan always chooses right, Chris bears full ex post causal responsibility for the unfair event ($r^{EP}_{Chris,unfair}(DD, rr) = 1$).[12] What if Chris *and* Dan always choose down? Then Chris is not pivotal anymore, but he would be if Dan would change his strategy. Thus, he bears partial ex post causal responsibility for the *unfair* event ($r^{EP}_{Chris,unfair}(DD, dd) = \frac{1}{2}$). The same holds for Dan. ▲

3.3 Ex ante causal responsibility

Two challenges arise when only ex post causal responsibility is used to evaluate causal responsibility. First, two agents can both be ex post causally responsible to the same degree even if, ex ante, their actions had vastly different probabilities of reaching that same degree. For example, compare two persons who each decide whether to shoot a gun or not. Person A has a 90 percent chance of killing an innocent bystander while person B has a 10 percent chance of killing an innocent bystander. Suppose both persons shoot and a bystander is killed in both cases. In this case, both persons would be attributed full ex post causal responsibility for the death of the bystander even if, ex ante, their actions had vastly different probabilities of reaching that level.

[12] Note that a single change of Chris' action is not enough to implement *fair*. However, distance from pivotality is measured by changes of strategies and, thus, Chris is evaluated as being pivotal for the unfair event even if he has to change two actions.

The ex ante causal responsibility component deals with such situations and allows to attribute higher causal responsibility to person A.

Second, it is often natural to assign responsibility to agents for events that they could have been pivotal for, even if these events are never implemented. Ex post causal responsibility is zero in this case. For example, suppose person A shoots but nature intervenes and the bystander is not killed. Ex ante causal responsibility captures that person A could have been pivotal for the death with some probability and should therefore be held responsible for that hypothetical event to some extent.

Ex ante causal responsibility incorporates these intuitions into the causal responsibility function. To this end, an agent i's ex ante causal responsibility for event x is defined like ex post causal responsibility with the sole difference that not the fully updated, but a partially updated behavior strategy profile is evaluated. In particular, the ex ante causal responsibility of agent i is evaluated at the history at which agent i is active for the first time on the path to terminal history z, $h_i(z)$.

Definition 5. *An agent i's degree of ex ante causal responsibility for any event $x \in X$ and any updated behavioral strategy profile $\sigma(h_i(z)) \in \Sigma(z)$ is defined as*

$$r_{i,x}^{EA}(\sigma(h_i(z))) = \mathbb{E}[\frac{1}{1 + d_{i,x}(s_i, \mathbf{s_{-i}})}|\sigma(h_i(z))] \tag{5}$$

where

$$\mathbb{E}[\frac{1}{1 + d_{i,x}(s_i, \mathbf{s_{-i}})}|\sigma(h_i(z))] = \sum_{\mathbf{s} \in S(h_i(z))} Pr(\mathbf{s}|\sigma(h_i(z))) \ \frac{1}{1 + d_{i,x}(s_i, \mathbf{s_{-i}})}.$$

Agents whose strategies induce a higher level of expected ex post causal responsibility bear a higher level of ex ante causal responsibility. A few features of ex ante causal responsibility are important to highlight: First, ex ante and ex post causal responsibility coincide when there are no moves of chance and when agents play pure strategies. Second, while ex post causal responsibility is only positive for the implemented event, ex ante causal responsibility allows to be responsible for multiple events, all for which a certain behavior strategy will potentially make one pivotal for.

Example 2 *(Sequential moves - cont'd).* Ex ante causal responsibility can also help to understand intuitions of responsibility attribution in sequential move games. In the example in which Alex can insult Bea, Bea can either be a stoic person who, when insulted, retaliates with 10% probability, or an easily irritable person who retaliates with 90% probability. If Alex insults stoic Bea, his ex ante causal responsibility for starting a fight is 0.1, but if he insults irritable Bea, his ex ante causal responsibility for starting a fight is 0.9. ▲

3.4 Overall causal responsibility

The two notions of ex ante and ex post causal responsibility are now combined as a convex combination to yield a function of overall causal responsibility.

Definition 6. *An agent i's degree of overall causal responsibility for event $x \in X$ is defined as*

$$r_{i,x}(\sigma(h_i(z)), \sigma(z)) = \gamma \cdot r^{EA}_{i,x}(\sigma(h_i(z)) + (1-\gamma) \cdot r^{EP}_{i,x}(\sigma(z)) \tag{6}$$

with $\gamma_R \in [0,1]$.

The parameter γ is an individual-specific parameter of the agent who evaluates the causal responsibility of agent i for event x and captures the weight that that individual places on ex ante vs ex post causal responsibility. An agent with $\gamma = 0$ places value on ex ante causal responsibility only and an agent with $\gamma = 1$ only considers ex post causal responsibility. When $\gamma \in (0,1)$, a combination of ex ante and ex post causal responsibility is used.[13]

Definitions 1 to 6 provide a guide for how to evaluate the causal responsibility of an agent: First, list all possible cases in which the agent would be pivotal. Then, calculate the distance between all these cases and the actual case under consideration. The minimum of these is the distance from pivotality. Finally, the distance from pivotality determines the ex post, the ex ante, and, taken together, the overall causal responsibility of the agent.

4 Experimental evidence

For itself, the notion of causal responsibility is a useful theoretic tool for assessing responsibility in multi-agent situation. Additionally, it can also be used to predict how people form responsibility perceptions in reality. To test whether the theory is successful in making correct comparative-statics prediction about the formation of responsibility perceptions in people, I conducted a laboratory experiment which I report in the following.

4.1 General setup

In the experiments, subjects were presented with a sequence of abstract, hypothetical scenarios. In each scenario, two players, Person A and Person B, interact in varying ways to either implement event "X" or not. After the description of the scenario, the actions of Person A and B, which always implemented event X, were revealed. Subjects were then always asked the same question: *How responsible is Person A for the implementation of event X?*. They selected one of the following four answers: "Not at all", "Little", "Medium", "Very". There were no indications that I was specifically interested in *causal* responsibility or that subjects should think about the underlying causal structure of a scenario.

In total four scenarios were tested. Each scenario tests a different prediction of the causal responsibility model. Scenario 1 and 2 respectively test the predictions of ex post and ex ante causal responsibility in a simultaneous-move environment. Similarly, Scenario 3 and 4 respectively test the predictions of ex post and ex ante causal responsibility in a sequential-move environment. For each scenario, subjects view the two variations on the same computer

[13] The combination of ex ante and ex post evaluations is reminiscent of similar approaches in other domains. For example, Saito (2013) and Cappelen et al. (2013) show experimentally that ex ante as well as ex post considerations matter for the evaluations of the fairness of income distributions.

screen and state their responsibility perceptions for each. The variations are designed to vary the predictions of the causal responsibility while leaving everything else constant.[14]

Answers were incentivized by the method developed by Krupka and Weber (2013). At the end of each session, one scenario was randomly selected. Subjects earned 6 Euro if their answer to that scenario corresponded to the most frequent answer in the session. Otherwise, they earn 0 Euro. Thus, subjects were incentivized to state what they believe is the most common perception of responsibility among a group of people. In addition, they always received the show-up fee of 4 Euro.

4.2 Procedures

In total, 99 participants took part in the experiment. Each subject participated in only one of a total of four sessions. All sessions took place at the Cologne Laboratory for Economic Research (CLER) at the University of Cologne in 2018. Participants were recruited from the CLER subject pool with the software "ORSEE" (Greiner, 2003). The experiments were computerized with the software "z-Tree" (Fischbacher, 2007). Before subjects entered the lab, they randomly drew a place card that specified at which computer terminal to sit. Subjects found paper copies of the general instructions at their assigned computer terminals. These instructions explained the rules of the experiment, the general decision situation (i.e., that subjects will have to evaluate the responsibility of two persons in hypothetical situations), and the monetary incentives. Subjects received detailed instructions for each specific scenario on the computer screen. The instructions included comprehension questions that had to be answered correctly before the experiment could begin. All instructions were read aloud to ensure common information regarding the instructions. Sessions lasted about 45 minutes and subjects earned, on average, 6.97 Euro including a show-up fee of 4 Euro.

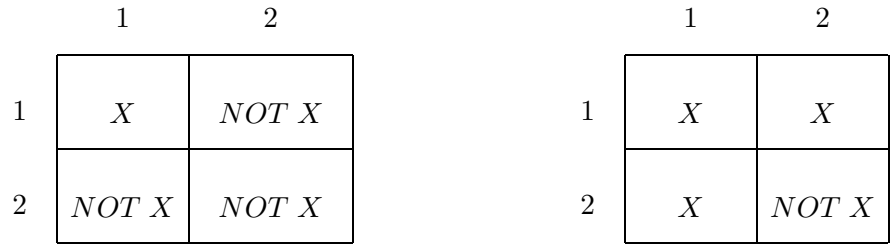
4.3 Scenarios and results

Scenario 1 *(Ex post - simultaneous moves).* The first scenario tests the main assumption underlying the causal responsibility model, namely that the distance from pivotality affects how responsible an agent is perceived to be for an event. Person A and Person B simultaneously choose between two options, Option 1 and Option 2. In Variation 1, the complements case, event X is only implemented if both persons choose Option 1, otherwise it is not implemented.

In Variation 2, the substitutes case, event X is implemented whenever at least one person chooses Option 1. Only if both persons choose Option 2, X is not implemented. For both variations, subjects learn that both persons simultaneously chose Option 1 and that X is therefore implemented. Thus, the two variations hold both persons' choices and the implemented event constant, and only vary whether Person A's choice is pivotal its implementation or not. They are then asked to rate Person A's responsibility for the implementation of X.

The notion of ex post causal responsibility rates Person A's responsibility higher in Variation 1 ($r^{EP}_{A,X}(1,1) = 1$) compared to Variation 2 ($r^{EP}_{A,X}(1,1) = 0.5$). Since no randomization by players was involved, ex ante and overall causal responsibility are identical to ex post causal responsibility. Thus, causal responsibility makes the following comparative-statics prediction:

[14] Table A.I in Appendix A.1 summarizes the scenarios and the number of participants.



(a) Variation 1 (b) Variation 2

Figure 3: Scenario 1

Prediction 1: Person A's responsibility is rated higher in Variation 1 than in Variation 2.

Table 1: Scenario 1 - Ex post - Simultaneous move (N=99)

Responsibility	"No"	"Little"	"Medium"	"Very"	p-value
Variation 1	2.02%	6.06%	34.34%	<u>57.58%</u>	0.004
Variation 2	6.06%	12.12%	<u>46.46%</u>	35.35%	

Notes: The percentages are the frequencies of the respective responses. The p-value is based on a Wilcoxon signrank test. The modal answers are underlined.

Table 1 summarizes the frequencies of the answers of the participants. The results show several interesting facts. First, subjects indeed rate Player A's responsibility significantly higher in Variation 1 compared to Variation 2 (Wilcoxon signrank test, $p = 0.004$). The modal answer of the subjects was that, Person A is "very" responsible in Variation 1 (stated by 57.58% of respondents), but only "medium" responsible in Variation 2 (stated by 46.46% of respondents). This confirms the prediction of the main assumption of the model, namely that distance from pivotality influences responsibility perceptions. It also means, that a pure "diffusion of responsibility model", which simply divides responsibility among all involved players and would, thus, predict equal responsibility in both Varations, does not correctly predict average responsibility ratings. Second, even though no player was pivotal for the event in Variation 2, only 6.06% of respondents said that Person A was not at all responsible for its implementation. Thus, not being pivotal does not shield a person from responsibility, as the causal responsibility model predicts.

Scenario 2 *(Ex ante - simultaneous moves).* The second scenario tests the assumption underlying ex ante causal responsibility, i.e., that responsibility also depends on the probability with which an agent will be pivotal for an event. Subjects are told that Person A and Person B simultaneously choose between two options, Option 1 and Option 2. This time, as in a coordination game, event X is only implemented if both persons choose the same option. Thus, whenever event X is implemented, both persons are always pivotal for it and bear full ex post causal responsibility. The variations of Scenario 2 differ in the probability with which Person B chooses Option 1. In Variation 1, Person B chooses Option 1 with 90% probability and in Variation 2, he chooses Option 1 with 10% probability. Subjects are then told that Person A and Person B indeed choose Option 1 and that event X is implemented. Again, subjects are then asked to evaluate Person A's responsibility for the implementation of event X.

	1 (90%)	2 (10%)
1	X	$NOT\ X$
2	$NOT\ X$	X

(a) Variation 1

	1 (10%)	2 (90%)
1	X	$NOT\ X$
2	$NOT\ X$	X

(b) Variation 2

Figure 4: Scenario 2

In both variations, ex post causal responsibility is always 1, while ex ante causal responsibility is higher in Variation 1 ($r^{EA}_{A,X}(1,(\frac{9}{10},\frac{1}{10})) = 0.9$) compared to Variation 2 ($r^{EA}_{A,X}(1,(\frac{1}{10},\frac{9}{10})) = 0.1$). If subjects place some weight on ex ante causal responsibility, overall causal responsibility of Person A for event X is also higher in Variation 1, and the theoretical prediction is thus:
Prediction 2: Person A's responsibility is rated higher in Variation 1 than in Variation 2.

Table 2: Scenario 2 - Ex ante - Simultaneous move (N=60)

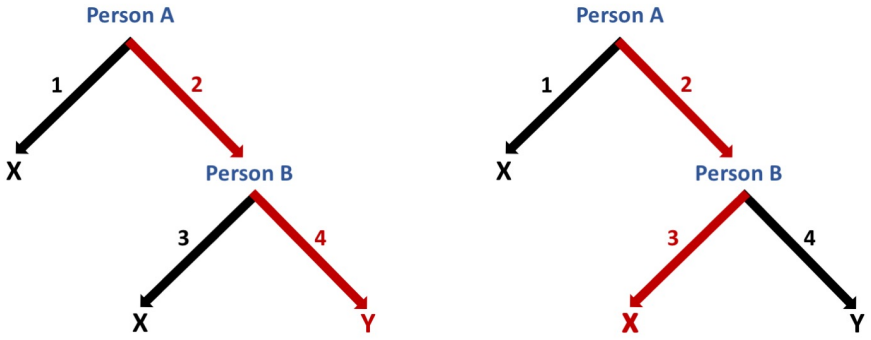
Responsibility	"No"	"Little"	"Medium"	"Very"	p-value
Variation 1	0.00%	1.67%	10.00%	<u>88.33%</u>	0.000
Variation 2	11.67%	<u>53.33%</u>	8.33%	26.67%	

Notes: The percentages are the frequencies of the respective responses. The p-value is based on a Wilcoxon signrank test. The modal answers are underlined.

Table 2 summarizes the responses of the subjects. Subjects attribute significantly higher responsibility to Player A in Variation 1 compared to Variation 2 (Wilcoxon signrank test, $p < 0.001$), which confirms Prediction 2. 88.33% of respondents state that Player A is "very" responsible in Variation 1, whereas only 26.67% do so in Variation 2. Thus, probabilistic reasoning, just as distance from pivotality, is crucial to explain variations in responsibility perceptions. Interestingly, in Variation 2, even though Player A choose the option (Option 1) that minimized the chance that event X would be implemented, only 11.67% of subjects state that Player A bears "no" responsibility for X. This is consistent with causal responsibility, which also predicts some responsibility even in this case.

Scenario 3 *(Ex post - sequential moves).* The third scenario tests ex post causal responsibility in a sequential game. In Scenario 3, Person A and Person B act sequentially, in a causal structure similar to that of Example 2. First, Person a decides between Option 1 and Option 2. If Person A chooses Option 1, event X is implemented immediately. If Person A chooses Option 2, Person B decides between Option 3 and Option 4. If he chooses Option 3, event X is again implemented. However, if he chooses Option 4, event Y is implemented. Thus, event Y is only implemented if Person A chooses Option 2, and, afterwards, Person B chooses Option 4.

Subjects are then again confronted with two variations. In both variations, Person A chooses Option 2. In Variation 1, Person B chooses Option 4 and in Variation 2 he chooses Option 3. Thus, Person A bears full ex post causal responsibility for event Y in Variation 1 (Person A is



(a) Variation 1 (b) Variation 2

Figure 5: Scenario 3 (N=39)

pivotal), and no ex post causal responsibility for event X in Variation 2 (Person A cannot be pivotal for event X after the choice of Option 2). Thus, the comparative statics predictions are that subjects' responsibility ratings for Person A should be higher in Variation 1.
Prediction 3: Person A's responsibility is rated higher in Variation 1 than in Variation 2.

Table 3: Scenario 3 (N=39)

Responsibility	"No"	"Little"	"Medium"	"Very"	p-value
Variation 1	5.13%	10.26%	<u>58.97%</u>	25.64%	0.000
Variation 2	25.64%	<u>35.90%</u>	33.33%	5.13%	

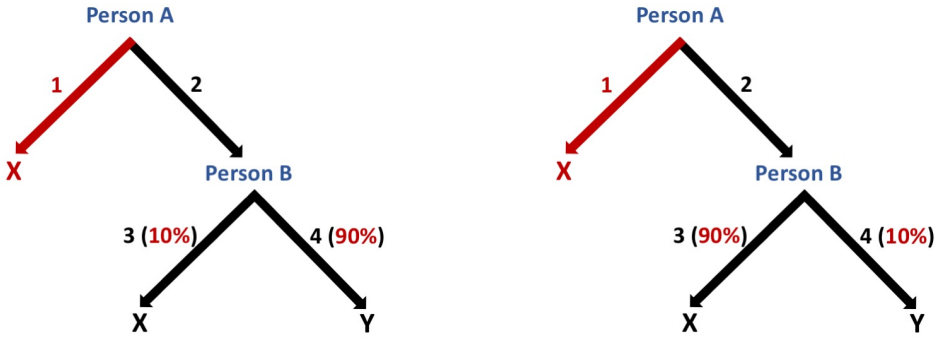
Notes: The percentages are the frequencies of the respective responses. The p-value is based on a Wilcoxon signrank test. The modal answers are underlined.

Table 3 summarizes the findings and shows that the prediction is confirmed. Responsibility ratings are significantly higher in Variation 1 (Wilcoxon signrank test, $p < 0.001$).
Scenario 4 *(Ex post - sequential moves 2).* The decision situation in Scenario 4 is identical to that in Scenario 3. However, subjects are now told that Person A chooses Option 1. Thus, event X is implemented and Person B doesn't move. The two variations of Scenario 4 vary with what probability Person would have choose Option 3 and 4, respectively, if he would have gotten to move. In Variation 1 and Variation 2, subjects are told that Person B will choose Option 4 with 90% or 10% probability, respectively. Thus, Scenario 4 tests whether potential moves in a part of the game tree that is never reached can affect responsibility ratings.

Subjects are asked to evaluate Person A's responsibility for the implementation of event X. If Person A chooses Option 1, she is pivotal for event X if Person B would choose Option 4, but she is one change away from being pivotal if Person B would choose Option 3. As the uncertainty regarding Person B's choice is never resolved, ex ante and ex post causal responsibility are equal in this case. Thus, overall causal responsibility is higher in Variation 1 ($r_{A,X}(1,(\frac{1}{10},\frac{9}{10})) = 0.95$) compared to Variation 2 ($r_{A,X}(1,(\frac{9}{10},\frac{1}{10})) = 0.55$) and the theory predicts:
Prediction 4: Person A's responsibility is rated higher in Variation 1 than in Variation 2.

As Table 4 shows, subjects' responsibility are significantly higher in Variation 1 (Wilcoxon signrank test, $p = 0.001$), which confirms the prediction.

The four scenarios showed that the causal responsibility notion is indeed capable of making correct comparative-statics predictions about peoples' actual and incentivized perceptions of



(a) Variation 1 (b) Variation 2

Figure 6: Scenario 4 (N=39)

Table 4: Scenario 4 (N=39)

Responsibility	"No"	"Little"	"Medium"	"Very"	p-value
Variation 1	0.00%	7.69%	5.13%	87.18%	0.001
Variation 2	0.00%	20.51%	38.46%	41.03%	

Notes: The percentages are the frequencies of the respective responses. The p-value is based on a Wilcoxon signrank test. The modal answers are underlined.

responsibility. If those responsibility perceptions influence people's preferences, the notion can be used to predict behavior, which I study in the next section.

5 Preferences for internal causal responsibility attribution

To study how responsibility perceptions influence behavior, I next incorporate the notion of causal responsibility in a preference framework. I start by assuming that people care about their own responsibility for the consequences of their actions, study the implications of such preferences for workers' effort provision, and test the behavioral predictions using existing experimental data.

5.1 Model

The setup is identical to the environment described in Section 3: A finite set of n agents and, potentially, nature take sequential and/or simultaneous actions, which together implement an event x. The agents' individual monetary payoff is determined by a function $\pi_i : Z \to \mathbb{R}$ that links the implemented terminal histories to payoffs. At the end of the game, the terminal history z, the implemented event x, and the resulting payoffs are common knowledge.

Each agent i potentially possesses *responsibility preferences*. Thus, he has a preference, in addition to his taste for monetary payoff, to seek causal responsibility for events that he judges as good and to avoid causal responsibility for events that he judges as bad. In the following, I introduce the components of a utility function that represents such preferences.

First, I define how agent i judges an event as good or bad. I assume that each event x generates a subjective payoff, $m_{i,x} \in \mathbb{R}$, for each agent i.[15] Each agent then evaluates the

[15]This payoff could be purely psychological, for example, when it is based on the moral evaluation of an event,

relative payoff of all possible events $x \in X$ according to a *judgment function* $j_{i,x}$. Agent i can judge an event x as *good* ($j_{i,x}(\cdot) > 0$), *bad* ($j_{i,x}(\cdot) < 0$), or *neutral* ($j_{i,x}(\cdot) = 0$). Specifically, I assume that the judgment of event x depends on the payoff it generates for agent i, $m_{i,x}$, relative to a reference payoff, $\bar{m}_i(X(h_i(s_i, \mathbf{s_{-i}})))$, that depends on the set of events that could possibly be implemented.

Definition 7. *For any agent $i \in I$ and pure strategy profile $(s_i, \mathbf{s_{-i}}) \in S$, agent i's judgment of event $x \in X(h_i(s_i, \mathbf{s_{-i}}))$ is given by the judgment function $j_{i,x} : X \times I \times S \to \mathbb{R}$ which is defined as*

$$j_{i,x}(h_i(s_i, \mathbf{s_{-i}})) = m_{i,x} - \bar{m}_i(X(h_i(s_i, \mathbf{s_{-i}}))) \tag{7}$$

with $\bar{m}_i(X(h_i(s_i, \mathbf{s_{-i}}))) = 0.5 \cdot [\min_{x \in X(h_i(s_i, \mathbf{s_{-i}}))} m_{i,x} + \max_{x \in X(h_i(s_i, \mathbf{s_{-i}}))} m_{i,x}]$.[16]

Similar to the restriction of an agent's potential causal responsibility to events that can realize after that agent's first move, I also restrict the reference payoff to only take these events into account, using again $h_i(s_i, \mathbf{s_{-i}}) \in H$, the history at which agent i moves first. $X(h_i(s_i, \mathbf{s_{-i}}))$ is then the set of events that could be implemented after the history at which agent i makes his first decision in strategy profile $(s_i, \mathbf{s_{-i}})$. Example 5 demonstrates the importance of this restriction.

Example 4 *(Relative judgment of events).* Figure 7 illustrates a sequential game. Assume that $m_{i,e_1} = m_{i,e_2} > m_{i,e_3}$ holds for both agents and terminal history $z = (in, r)$ is implemented. After this history, both agents are fully causally responsible for event e_2. However, when agent 1 made his choice, event e_3 was still possible, but not anymore when agent 2 chose. Therefore, the implementation of event e_2 is judged positively by agent 1, $j_{1,e_2}(\varnothing) > 0$, but neutrally by agent 2, $j_{2,e_2}(in) = 0$. ▲

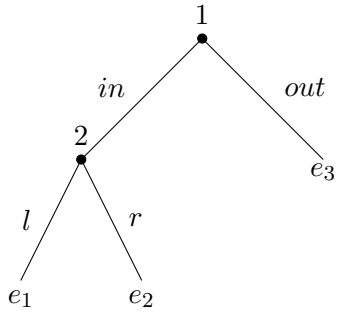


Figure 7: Relative judgment of events.

Formulating the judgment function in relative terms has three appealing features. First, when only two events are possible, they are judged neutrally only when both generate the same payoff, i.e., in case of indifference. This captures the fact that people don't care about responsibility for two equally "good" or "bad" events, when nothing else was possible. Second, it provides an intuitive scale: The more an event's payoff deviates from the reference payoff, the better or worse it is judged to be. And, hence, the more or less does agent i care about his

or purely monetary, for example, when it is based on the success or failure of a work project.

[16] The formulation of the reference payoff is thus similar to that of intention-based social preference models (c.f. Rabin, 1993; Dufwenberg and Kirchsteiger, 2004).

responsibility for that event. Third, since events are judged relative to the set of events that a specific agent could be responsible for, it is possible that the same event is judged differently by different agents.

Next, agent i evaluates his causal responsibility for each possible event, which depends on his and all other agents' strategies. Hence, agent i has to form first-order beliefs about the other agents' strategies. Agent i's initial first-order belief about agent j's behavior strategy is denoted by $\alpha_{ij} \in \Sigma_j$ and the updated belief at history h by $\alpha_{ij}(h)$. Beliefs are updated according to Bayes' rule. Histories become public information as soon as they occur, which implies that the probability of any action that is actually observed is replaced by 1 whereas probabilities of actions in other part of the game tree are kept intact, as described before. Agent i's belief about moves of nature is always accurate, $\alpha_{ic} = \sigma_c$.

Note that, in Section 3, we evaluated ex ante and ex post causal responsibility separately, as we assumed knowledge of the terminal history of the game. However, an agent who evaluates his own future causal responsibility when choosing his strategy is potentially still uncertain about the relevant future choices of others. Therefore, he will choose as if only ex ante causal responsibility is relevant (i.e., $\gamma_i = 1$).[17] Agent i's causal responsibility for event x is then given by the function, $r_{i,x}(\sigma_i, (\alpha_{ij})_{j \in I_c})$, and his utility function is defined as:

Definition 8. *The expected utility of agent i is a function $u_i : Z \times \Sigma \to \mathbb{R}$ that is defined as*

$$U_i(\sigma_i, (\alpha_{ij})_{j \in I_c}) = \pi_i(\sigma_i, (\alpha_{ij})_{j \in I_c}) + \rho_i \sum_{x \in X} [r_{i,x}(\sigma_i, (\alpha_{ij})_{j \in I_c}) \cdot j_{i,x}(h_i(s_i, \mathbf{s_{-i}}))]. \tag{8}$$

The parameter $\rho_i \geq 0$ captures how much agent i cares his responsibility for the events compared to his own monetary payoff. A utility-maximizing agent with responsibility preferences (i.e., $\rho_i > 0$) will seek to reduce his responsibility for events that he judges as bad and increase his responsibility for events that he judges as good.

The game is fully specified as $\Gamma = (I_c, \Sigma, (U_i)_{i \in I})$ and the equilibrium can be defined. I consider a complete information framework in which the rules of the game and agents' preferences are common knowledge. Since beliefs about other agents' strategies directly enter each agent's utility function, the game is a dynamic psychological game and I therefore apply an equilibrium concept similar to that developed in Dufwenberg and Kirchsteiger (2004).

Definition 9. *The profile $\sigma^* = (\sigma_i^*)_{i \in I}$ is a* sequential responsibility equilibrium *if it holds that*

1. $\sigma_i^*(h) \in \arg\max_{\sigma_i \in \Sigma_i(h_i(s_i, \mathbf{s_{-i}}))} U_i(\sigma_i, (\alpha_{ij})_{j \in I_c})$ *for all $i \in I$ and $h \in H$, $z \in Z$*
2. $\alpha_{ji} = \sigma_i \; \forall \; i, j \in I$.

Condition 1 stipulates that the *sequential responsibility equilibrium* is a strategy profile such that at each history $h \in H$ all agents take actions that maximize their utility given their beliefs and given that they follow their equilibrium strategy in other histories. Condition 2 says that, in equilibrium, initial beliefs are correct. At any subsequent history, beliefs assign probability one to the sequence of choices that define that history, but are otherwise identical to the initial beliefs.

[17] In Section 6, in which I study external causal responsibility attribution, this restriction will not hold.

5.2 Application: The incentive effects of internal causal responsibility attribution

To demonstrate the effects of internal causal responsibility attribution, I examine how it influences incentives for effort provision. Specifically, I study a situation in which either one worker or a team of workers has to finish a project before a deadline. As an example, one could think of an editor who sends a paper for review to one or more referees. I analyze whether it matters for the timing of effort provision whether the workers do or do not feel causally responsible for the timely completion of the project. Furthermore, study how the number of workers and the distribution of their preferences matters.

The specific setting is the following: a manager delegates a task to one or multiple workers and sets a deadline of two days for the completion of the whole project. Each worker needs one day to finish her task and can, thus, decide whether to work on day one, $a_i = w$, and enjoy leisure on day two, or to shirk on day one, $a_i = s$, and work on day two. Working comes with an effort cost, $c(w) < 0$, while shirking provides some some utility from leisure, $c(s) > 0$. The project is completed when all workers have finished their task. The relevant events are therefore whether the project is completed on day one or on day two, which I call, respectively, *early* or *late* completion, $X = \{e, l\}$. For the company, it is beneficial if the task is completed as soon as possible. I assume that the workers internalize some of the benefits and costs of early or late completion and, therefore, judge early completion positively, $j_{i,E} > 0$, and late completion negatively, $j_{i,L} < 0$. Thus, the workers face a trade-off between the benefits from leisure and the costs of working and their causal responsibility for a positive or negative event. I assume that the workers realize any utility from their causal responsibility for the events on day one since this is the date at which the event is determined. The overall discounted utility of a worker i on day one is therefore:

$$U_i(a_i, a_{-i}) = \begin{cases} c(s) + \rho_i \, r_{i,L}(s, a_{-i}) \, j_{i,L} + \beta \, c(w) & \text{if } a_i = s \\ c(w) + \rho_i \, r_{i,E}(w, a_{-i}) \, j_{i,E} + \beta \, c(s) & \text{if } a_i = w \end{cases}$$

where $\beta \in (0, 1)$ is the worker's discount factor. A worker will finish the task on day one, if

$$\rho_i > \frac{(1-\beta) \, (c(s) - c(w))}{r_{i,E}(w, a_{-i}) \, j_{i,E} - r_{i,L}(s, a_{-i}) \, j_{i,L}}. \tag{9}$$

Single worker case: First, I consider that only a single worker is tasked to complete the project. If that worker possesses standard preferences without any concern for her causal responsibility ($\rho_i = 0$), then the worker would always shirk on day one and complete the project on day two. If, on the other hand, the worker cares enough about her causal responsibility for the completion or the delay of the project ($\rho_i > 0$), she will finish the task on day one, if

$$\rho_i > \frac{(1-\beta) \, (c(s) - c(w))}{j_{i,E} - j_{i,L}} = \bar{\rho}. \tag{10}$$

That is because the worker is always pivotal and thus fully causally responsible for the respective events ($r_{i,L}(w) = r_{i,L}(s) = 1$). Thus, internal causal responsibility attribution acts as an

incentive to provide effort early and can explain why a worker might finish a task before the actual deadline.

Homogeneous workers, changing team size: Next, I study how the worker's decision changes, when the completion of the project requires the input of several workers. First, I assume that workers are homogeneous, i.e., they care to the same extent about their causal responsibility for the project, and study how a change in the number of workers affects the potential equilibria of the game. Note that if all workers have standard preferences of pure monetary self-interest, the problem is still one of individual decision-making—the workers' actions don't influence each other's utility—and all workers would shirk on day one. However, if the workers care about their causal responsibility for the project, the problem is one of strategic decision-making as each worker's decision potentially influences all other workers' causal responsibility for the events.[18] Since one shirking worker is enough to delay the project, the workers are perfect complements for the early and perfect substitutes for the late completion of the project. Therefore, when all workers work on day one, each of them is fully causally responsible for the early completion and one deviating worker would be fully causally responsible for the late completion. Hence, all working on day one exists as an equilibrium, if each worker fulfills Condition 10, just as in the single-worker case. The existence of this equilibrium is therefore independent of the team size.

However, there also exists an equilibrium in which all workers shirk. If all shirk, each workers' causal responsibility for the late completion of the project decreases with the team size n ($r_{i,L}(s,...,s) = \frac{1}{n}$). No worker would deviate and work, if their regard for causal responsibility satisfies

$$\rho_i < \frac{(1-\beta)\,(c(w)-c(s))}{r_{i,L}(s,...,s)\;j_{i,L}} = \frac{n\,(1-\beta)\,(c(w)-c(s))}{j_{i,L}} = \rho(n).$$

Since $\frac{\partial \rho(n)}{\partial n} > 0$, the existence of an equilibrium in which all shirk depends on the number of team members: the larger the team, the larger the interval for which it exists.

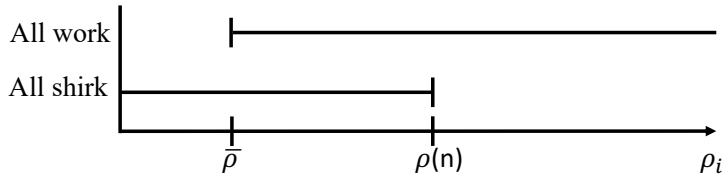


Figure 8: Equilibria with homogeneous workers and changing team size.

Figure 8 shows the interval of ρ_i for which the "all-work" and the "all-shirk" equilibria exists. There are three cases. First, when the workers' concern for causal responsibility is small (i.e., $\rho_i < \bar{\rho}$), all shirking is the unique equilibrium. Second, all working exists as the unique equilibrium, if the workers' concern for causal responsibility is high (i.e., $\rho_i > \rho(n)$). However, the interval in which all working is the unique equilibrium decreases with an increase in the team size. Finally, for intermediate concern for causal responsibility (i.e., $\bar{\rho} \leq \rho_i \geq \rho(n)$) both equilibria exist and the team members can coordinate on either of them. Several equilibrium-selection criteria have been proposed to predict which equilibrium will be selected in coordination games. Most prominently, Schelling (1980) proposed that, if equilibria can be Pareto-ranked, the payoff-dominant equilibrium will be coordinated on. In our case, the payoffs of the "all-work"

[18] In the interest of brevity, I focus on pure strategy equilibria.

equilibrium are independent of the team size, while the payoffs of the "all-shirk" equilibrium increase with the team size. Hence, while the "all-work" equilibrium is payoff-dominant for small team sizes, the "all-shirk" equilibrium can become payoff-dominant for larger team sizes.[19]

To summarize, compared to the single-worker case, needing a team of workers makes it more likely that the completion of the project is delayed as the causal responsibility for the delay is more and more diffused the larger the team is. For the example from above, the implications are that if multiple referees are asked to write a report and they know that the review process is only finished after all hand in their reports, then they feel less responsible for the editor's and the authors' waiting time compared to when they are solely responsible and thus might delay working on the report more readily.

Heterogeneous workers, constant team size: Finally, I study the implications of heterogeneity in the team members' concern for causal responsibility when the team size remains constant. Specifically, I assume that there is a fixed number of four workers. As before, an equilibrium with early completion requires that all four workers concern for causal responsibility satisfies Condition 10. A single worker whose concern for causal responsibility is low enough to not satisfy the condition will shirk and thereby delay the project.

But when do workers start to shirk that would not shirk when alone, i.e., who do not violate Condition 10? As we have seen before, increasing the number of workers can lead to an increase in shirking, as causal responsibility gets diffused among more workers. However, the same can happen with a constant number of workers but a change in the distribution of ρ_i within the workers. To classify subjects, I let ρ^n denote the concern for causal responsibility of a worker who would shirk, if at least n workers shirk in total, but not if $n-1$ workers shirk.[20] A worker with ρ^1 would always shirk independent of how many others shirk, a worker with ρ^3 would shirk if at least two other workers also shirk, and so on. In a team of four workers, a worker with ρ^5 would never shirk, independent of what the others do.

Table 5: Preference distribution in three cases.

	ρ^1	ρ^2	ρ^3	ρ^4	ρ^5
Case 1	-	2	-	2	-
Case 2	-	1	1	1	1
Case 3	1	1	1	1	-

Table 5 shows three cases with different distributions of the four workers' concern for causal responsibility. In Case 1, two workers would shirk if one other worker would also shirk, while two workers would only shirk if three others would also shirk. Hence, in this case, there are three equilibria. One in which all workers work and the project is completed early, one in which all shirk and the project is completed late, and a third one in which the two workers with low concern for causal responsibility shirk and the two workers with high concern for causal responsibility do not shirk.

In Case 2, all four workers fall into different categories regarding their concern for causal responsibility. In particular, there is one worker who will shirk if one other worker shirks, one

[19] More precisely, if $\rho_i < \frac{(1-\beta)\ (c(s)-c(w))}{j_{i,E}}$, then the "all-shirk" equilibrium will become payoff-dominant for large enough team sizes.

[20] Formally, $\rho^n \in [\frac{(n-1)\ (1-\beta_i)\ (c(w)-c(s))}{j_{i,L}}, \frac{n\ (1-\beta_i)\ (c(w)-c(s))}{j_{i,L}}]$.

worker who will shirk if two other workers shirk, one worker who will shirk if three other workers shirk, and one very responsible worker who will never shirk, independent of what the others do. In this case, the *unique* equilibrium is one in which all work on day one and the project is completed early. This is, because none of the workers who would shirk conditional on others shirking finds enough "allies" to reduce causal responsibility enough to make shirking attractive for each of them.

In Case 3, on the other hand, three of the four workers are identical to Case 2, but the one worker who never shirks is replaced by a worker who always shirks. This single replacement reverses the previous outcome. Now, the *unique* equilibrium is one in which all workers shirk and the project is completed late. This change occurs, because the single "irresponsible" worker triggers a domino effect that leads all workers to shirk. In particular, knowing that one worker will always shirk, shirking becomes attractive for the worker with ρ^2. This, in turn, is known by the worker with ρ^3, and so on, until also the last, highly responsible worker with ρ^4 decides to shirk.

To summarize, whereas in Case 2 a single conscientious individual induced all other workers to work, a single carefree worker induced all others to shirk in Case 3. Thus, these cases show that causal responsibility considerations can lead to an "one-bad-apple-spoils-the-barrel" effect as well as the opposite. They highlight the destructive nature of *irresponsible* workers as well as the motivating, role-model type nature of *responsible* workers.

Comparison to other theories. Can other theories predict the same behavioral patterns? Standard preferences of pure monetary self-interest predict that all workers shirk and the project is completed late. However, causal responsibility theory is only one among many theories that incorporate non-monetary arguments in the utility function. Therefore, it is instructive to compare the predictions of causal responsibility theory to that of other theories, in particular, theories of social preferences.

Theories based on outcome-based social preferences, such as inequity aversion (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), argue that people, in addition to their preference for monetary payoff, care about the distribution of monetary payoffs between themselves and others and prefer, ceteris paribus, to reduce the inequity between themselves and others. A strict application of the theory does not predict any difference in behavior compared to the standard model, as all workers receive the same monetary payment at the end of day two. However, a looser application of the theory could incorporate that the workers also dislike utility differences among themselves. If that is the case, inequity aversion predicts that all workers shirking on day one is always an equilibrium. In that scenario, payoff is maximized for all workers and no inequality exists. In addition, there might also exist an equilibrium in which all workers work on day one. No one would deviate, if the benefit from shirking are not enough to outweigh the disutility from the resulting inequity among the shirking and the working workers. How would group size impact the predicted equilibria? First, in groups of all sizes shirking on day one is an equilibrium as it again provides the highest monetary payoff with no payoff inequality to all workers. However, group size can impact whether an equilibrium with all workers working exists. In larger groups, deviating from such an equilibrium through shirking is more costly as it puts the shirking worker ahead of more other workers who work. Hence, the larger the group,

the more likely it is that an all-working equilibrium exists. This is the opposite of what causal responsibility predicts.

Another social preference theory incorporates that people care, in addition to monetary payoff, also about intentions (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004). In particular, the theory predicts that people act reciprocally: they reward kind behavior of others towards themselves and punish unkind behavior. However, in the application, (un)kindness does not exist as the workers' payoffs are not influenced by each other's actions. Hence, the predictions of intention-based social preferences coincide with those of the standard model.

Finally, Bartling and Fischbacher (2012) also formulate a theoretic notion of responsibility which is therefore a natural comparison. However, since their notion of responsibility is not incorporated in a utility function, it is not exactly suitable to make game-theoretic predictions. In their theory, responsibility is divided among those players whose action increase the likelihood of an event, relative to an ex ante belief about the behavior of the players. If we assume that team members are ex ante expected to act according to the two equilibria that were identified before—all-work or all-shirk—then the responsibility notion of Bartling and Fischbacher (2012) makes the following predictions: First, if all work, none of the workers bears responsibility for the early completion because none increased its likelihood relative to the ex ante beliefs even though they all are pivotal. However, if one worker deviates and shirks, he bears full responsibility for the late completion as she was the single worker that increased its likelihood. Incorporated in a utility framework, this could deter her from shirking and, thus, sustain the equilibrium. Second, if all shirk, again, none of the workers bears any responsibility for the late completion because none increased its likelihood relative to the ex ante beliefs. Since no one is responsible, no worker can redeem herself from responsibility by deviating and working. Hence, since working is costly, deviation has no benefits, and therefore all shirking is always an equilibrium. Importantly, the existence of both equilibria is independent of the number of workers. Thus, the predictions of this notion of responsibility differ substantially from mine.

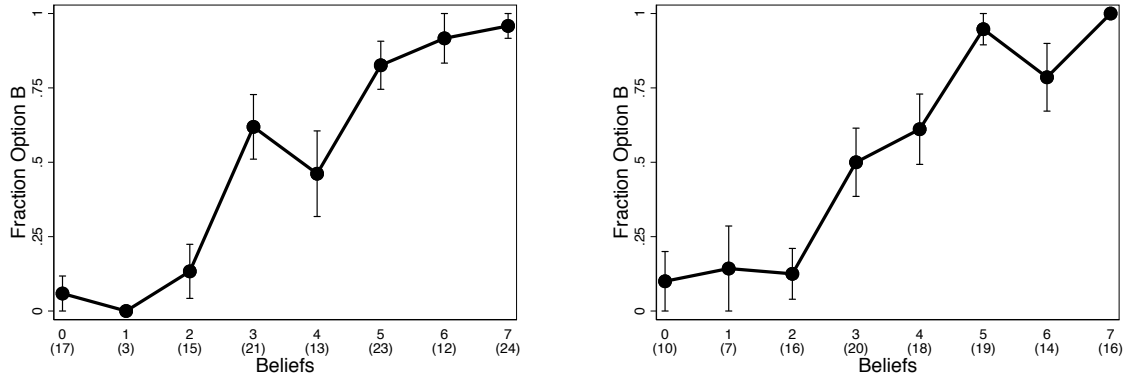
5.3 Discussion of related experimental findings

As demonstrated in Section 4, the notion of causal responsibility successfully predicts incentivized responsibility perceptions in laboratory experimental settings. In the following, I examine if it can also successfully predict behavior that, according to the theory, should be influenced by causal responsibility considerations.[21]

The most direct evidence for the role of causal responsibility considerations in group decision-making stems from a study on diffused pivotality by Falk, Neuber, and Szech (2020). In their

[21]There exists a large experimental literature comparing individual to group moral behavior that generally finds that people are more likely to act selfishly when part of a group than when acting alone (e.g., Cason and Mui, 1997; Dana, Weber, and Kuang, 2007; Luhan, Kocher, and Sutter, 2009; Behnk, Hao, and Reuben, 2017; Kocher, Schudy, and Spantig, 2018). This effect has sometimes been attributed to a diffusion of responsibility in groups. However, comparisons between individual and group choices are often not suitable for a clean test of causal responsibility theory, as they introduce changes to the decision environment other than changes in causal responsibility. For example, the study of Dana, Weber, and Kuang (2007) compares dictator games with single and multiple dictators and finds that subjects are more selfish when acting in groups. This result cannot be explained by causal responsibility theory as dictators are similarly causally responsible in both settings. However, in the multiple dictator compared to the single dictator treatment, the selfish outcome benefits multiple dictators while the negative externality stays the same. Thus, the selfish outcome might be judged less badly and this could have led to the increased adoption of it despite the fact that causal responsibility levels remain unchanged.

"diffused pivotality" treatments, participants are matched in groups of eight and individually choose between two options. Choosing Option B grants participants a higher monetary payoff than Option A. However, Option B also implements a negative externality if it is chosen by *at least one* out of the eight group members. Depending on the treatment, the negative externality is either the death of eight mice or the substantial reduction of a donation to a charity. Thus, while monetary payoffs depend on each participant's individual choice, causal responsibility for the negative externality depends on the decisions of all group members. Participants who care only about monetary payoff will choose Option B independent of their beliefs about the choices of the other group members. On the other hand, participants who, in addition, care about their causal responsibility for the implementation of the negative externality, will react to their beliefs about the choices of the other group members. In particular, Option B becomes relatively more attractive the more other group members they believe will choose Option B. That is, because causal responsibility for the negative externality and the associated disutility decreases accordingly. If causal responsibility is indeed a driver for choices, then participants' likelihood of choosing Option B should increase the more other group members a participant believes will also choose Option B.



(a) Mouse treatment.

(b) Charity treatment (round 1).

Figure 9: Share of subjects in the respective treatment choosing Option B depending on their belief about the number of other group members choosing Option B.
Notes: Graphs are generated using the published data of FNS (2020). Error bars show standard errors. Numbers of observations are shown in brackets.

The study by Falk, Neuber, and Szech (2020) elicits incentivized beliefs regarding the participants expectations about the number of other group members who choose Option B. This allows a direct statistical test of the theory. Figure 9 shows the share of subjects in the respective treatment choosing Option B depending on their belief about the number of other group members choosing Option B. As is clearly visible, there is a positive and strongly significant correlation in the predicted direction between the beliefs and the choices of participants (Spearman rank-order correlation coefficient: Mouse: $\rho = 0.654$, $p < 0.001$; Charity: $\rho = 0.634$, $p < 0.001$). Furthermore, the figures also demonstrate that it is distance from pivotality that is driving the effect and not whether a participant believes that he or she is pivotal or not. If the latter would be the case, then we would expect a jump in the likelihood of choosing Option B between beliefs

of zero and one, and no further increase thereafter. This is not the case.[22][23]

6 Preferences for external causal responsibility attribution

Next, to also study how causal responsibility perceptions influence how people react to the actions of others, I incorporate them in a framework of external causal responsibility attribution. The framework is then used to study the implications for punishment and reward behavior. The basic setup is a two-stage game consisting of a *collective action stage* and a *responsibility attribution stage*.

6.1 Model

Stage 1 - Collective action stage. The setup of stage 1 is identical to the environment described in Section 3: A group of n stage-1 agents and, potentially, nature take sequential and/or simultaneous actions which implement an event. Their individual monetary stage-1 payoff is determined by a function $\pi_i : Z \to \mathbb{R}$. In addition to the stage-1 agents, there exists an agent $R \notin I$ who is inactive in stage 1. Agent R cares about the possible events, which is captured by his payoff function $m_{R,x} : X \to \mathbb{R}$. At the end of stage 1, the terminal history z, the implemented event, and the resulting stage-1 payoffs are common knowledge.

As an intuitive example, the stage-1 agents could be thought of as a set of firms, each of which decides whether to produce with a dirty and cheap or a clean and expensive technology and which only care about maximizing profits. In this situation, agent R could be thought of as a representative consumer, who cares about the destruction of the environment and will hold firms responsible for it.

Stage 2 - Responsibility attribution stage. In stage 2, agent R judges the possible stage-1 events and the stage-1 agents' causal responsibility for them. He is assumed to possess *responsibility preferences*. Thus, he has a preference, in addition to his taste for monetary payoff, to reward or punish stage-1 agents for the implementation of what he judges as good or bad events, but only to the extent that those agents are causally responsible for them. The components of a utility function that represents such preferences are similar to Section 5, but with a few modifications.

[22] Note that internal causal responsibility does not necessarily imply in the probability of choosing Option B when beliefs increase from zero to one. That is, because, while responsibility decreases, the overall judgment can still be negative and only become positive for larger distances from pivotality.

[23] The discussed treatments have the advantage that beliefs about distance from pivotality vary but everything else is held constant. Additionally, the study of FNS also includes a baseline in which each subject individually chooses between Option A and Option B. While the monetary payoffs are identical, Option B now implements the death of one mouse or the smaller reduction of a donation to a charity for each participant individually. Thus, in this baseline, every participant knew that they are fully causally responsible for the relatively smaller externality. Causal responsibility theory suggests that subjects feel, on average, more responsible in this baseline treatment than in the group treatment, as in the latter the average distance from pivotality is increased and, thus, participants should be more likely to choose Option B. Indeed, this is what the study finds (two-sample test of proportions, Mouse: $p = 0.04$, Charity: $p = 0.004$). Thus, also this comparison is in line with the predictions of causal responsibility theory. However, comparisons between individual and group choices are often not suitable for a clean test of causal responsibility theory, as they introduce changes to the decision environment other than changes in causal responsibility. In this case, a pivotal choice leads to one dead mouse / a donation reduction of 15 Euro in the baseline treatment, but to the to eight dead mice / a donation reduction of 120 Euro in the group setting. Thus, it is not surprising that pivotal subjects in the baseline are significantly more likely to choose Option B than those who believe they are pivotal in the group treatments: the externality is imply much larger.

First, agent R evaluates the stage-1 events $x \in X$ according to the following *judgment function* $j_{R,x}$:

Definition 10. *For any agent $i \in I$ and history $z \in Z$, agent R's judgment of event $x \in X(h_i(z))$ is given by the judgment function $j_{R,x} : X \times I \times Z \to \mathbb{R}$ which is defined as*

$$j_{R,x}(h_i(z)) = m_{R,x} - \bar{m}_{R,x}(X(h_i(z))) \tag{11}$$

with $\bar{m}_{R,x}(X(h_i(z))) = 0.5 \cdot [\min_{x \in X(h_i(z))} m_{R,x} + \max_{x \in X(h_i(z))} m_{R,x}]$.

$X(h_i(z))$ is the set of events that could be implemented after the history at which agent i makes his first decision on the path to terminal history z. Thus, agent R's judgment of a single event can still differ depending on which agent i he is evaluating.

Second, agent R evaluates the causal responsibility of each stage-1 agent for each possible event. Agent R's initial first-order beliefs about agent i's behavior strategy is denoted by $\alpha_{Ri} \in \Sigma_i$ and the updated beliefs by $\alpha_{Ri}(h)$. Beliefs are updated as before. Agent i's causal responsibility for event x, as evaluated by agent R, is therefore given by the function, $r_{i,x}((\alpha_{Ri}(h_i(z), \alpha_{Ri}(z))_{i \in I_c})$.

Given how agent R judges the stage-1 events and evaluates the agents' causal responsibility for them, agent R's *overall judgment of the behavior of agent i* comprises the sum over the judgments of all possible events, weighted by agent i's causal responsibility for them, $\sum_{x \in X(\cdot)} r_{i,x}(\cdot) \cdot j_x(\cdot)$. Similar to the judgment of events, behavior can be judged as *praiseworthy* ($\sum_{x \in X(\cdot)} r_{i,x}(\cdot) \cdot j_{R,x}(\cdot) > 0$), *blameworthy* ($\sum_{x \in X(\cdot)} r_{i,x}(\cdot) \cdot j_{R,x}(\cdot) < 0$), or *neutral* ($\sum_{x \in X(h_i(z))} r_{i,x}(\cdot) \cdot j_{R,x}(\cdot) = 0$).

The overall judgment of agent i's behavior triggers a reaction by agent R. Specifically, agent R can choose an allocation $p_i \in P_i(z)$ for each stage-1 agent, where $P_i(z)$ is the set of feasible allocations for agent i after history z and $P(z) = \prod_{i \in I} P_i(z)$ is agent R's action space after history z. A behavior strategy for agent R is a function $\sigma_R \in \Sigma_R$ that associates with every history $z \in Z$ a probability distribution over $P(z)$. Agent R is said to *punish* agent i if he reduces his stage-2 payoff ($p_i < 0$), and he is said to *reward* agent i if he increases his stage-2 payoff ($p_i > 0$).[24]

Definition 11. *The expected utility of agent R is a function $U_R : Z \times \Sigma \times \Sigma_R \to \mathbb{R}$ that is defined as*

$$\begin{aligned} U_R(\sigma_R(z), (\alpha_{Ri}(h_i(z)), \alpha_{Ri}(z))_{i \in I_c}) &= \pi_R(\sigma_R(z)) \\ &+ \rho_R \sum_{i \in I} \left[\sum_{x \in X(h_i(z))} r_{i,x}((\alpha_{Ri}(h_i(z)), \alpha_{Ri}(z))_{i \in I_c}) \cdot j_{R,x}(h_i(z)) \right] \cdot p_i(\sigma_R(z)). \end{aligned} \tag{12}$$

The parameter $\rho_R \geq 0$ captures how much agent R cares about punishing or rewarding the behavior of the stage-1 agents compared to his own monetary payoff. If agent R cares about attributing responsibility to some extent, i.e., $\rho_R > 0$, then he will match the signs of his overall judgment of behavior of agent i and the allocation to agent i in order to maximize his utility.

[24] The implicit reference allocation is therefore zero, the allocation that neither increases nor decreases agent i's payoff. This is the natural allocation reference when thinking about punishment and reward.

Hence, an overall blameworthy (praiseworthy) behavior of agent i triggers punishment (reward) of agent i by agent R.[25]

The function $U_i : \Sigma \times \Sigma_R \to \mathbb{R}$ denotes the expected overall utility of agent i from the game as a whole which is simply the expected sum of his monetary payoffs in the two stages. The stage-1 agents choose strategies to maximize their expected utility, rationally anticipating the behavior of agent R in stage 2. To do this, they have to form second-order beliefs about R's beliefs about all stage-1 agents' strategies, denoted by $\beta_{iRj} \in \Sigma_j \ \forall \ j \in I_c$. Given these initial beliefs, agents update their beliefs, knowing that agent R will observe the same history, up until they have to make their own move. The expected overall utility of agent R is simply the expected sum of his utility from stage 1 and 2, $U_R : Z \times \Sigma \times \Sigma_R \to \mathbb{R}$.

The game is thus fully specified as $\Gamma = (I_c \cup \{R\}, \Sigma, \Sigma_R, (U_i)_{i \in I}, U_R)$ and the equilibrium can be defined. I consider a complete information framework in which the rules of the game and agents' preferences are common knowledge. Since beliefs about other agents' strategies directly enter agent R's utility function, the game is a dynamic psychological game and I, therefore, apply an equilibrium concept similar to that developed in Dufwenberg and Kirchsteiger (2004).

Definition 12. *The profile* $\sigma^* = (\sigma_i^*)_{i \in I}$ *and* σ_R^* *is a* sequential responsibility equilibrium *if it holds that*

1. $\sigma_i^*(h) \in \arg\max_{\sigma_i \in \Sigma_i(h_i(z))} U_i(\sigma(h_i), \beta_{iRj}(h_j(z))_{j \in I_c})$ *for all* $i \in I$ *and* $h \in H$, $z \in Z$
2. $\sigma_R^*(z) \in \arg\max_{\sigma_R \in \Sigma_R} U_R(\sigma_R(z), (\alpha_{Ri}(h_i(z)), \alpha_{Ri}(z))_{i \in I_c})$ *for all* $z \in Z$
3. $\beta_{iRj} = \alpha_{Rj} = \sigma_j \ \forall \ i, j \in I$.

Condition 1 stipulates that the *sequential responsibility equilibrium* is a strategy profile such that at each $h \in H$ all stage-1 agents take actions that maximize their utility given their beliefs and given that they follow their equilibrium strategy in other histories. Similarly, condition 2 stipulates that at each history $z \in Z$ agent R makes allocation choices that maximize his utility given his beliefs. Condition 3 says that, in equilibrium, initial beliefs are correct. At any subsequent history, beliefs assign probability one to the sequence of choices that define that history, but are otherwise as the initial beliefs.

6.2 Application: Designing voting rules

In the following, the game structure that was set up in Section 6.1 is used to analyze how causal responsibility attribution can affect voting outcomes under different voting rules compared to standard preferences. The *collective-action stage* consists of a simultaneous move of n politicians who vote on whether to implement a reform or not, $a_i \in \{y, n\}$. The $Reform$ is only implemented if at least t politicians vote for it. The true state is that the reform is *needed* and all politicians are aware of that, but not the public. For example, politicians might have information that an increase in the retirement age is needed due to demographic changes. However, politicians get a private, *expressive* payoff from voting *no*, $\pi_i(n) > \pi_i(y)$.

[25] I assume that agent R's punishment and reward is symmetrical, i.e., responsibility for a bad event is punished to the same extent as responsibility for a equally sized good event is rewarded. Recent experimental supports this assumption (Anselm et al., 2022).

Agent R, in this case, represents the public. The public's payoff at the end of the *collective-action stage* is strictly greater when the reform is implemented ($m_{R,reform} > m_{R,no\,reform}$). In a second stage, the public gets the chance to punish or reward the politicians for their voting behavior, e.g., by attributing votes in the following election. Formally, agent R chooses an allocation for each politician i, $p_i \in [\underline{p}, \overline{p}]$ with $\overline{p} > 0 > \underline{p}$. Agent R faces convex allocation costs such that his monetary stage-2 payoff equals $\pi_R(\mathbf{p}) = -\frac{c}{2}\sum_{i\in I} p_i^2$, with $c > 0$.

Before discussing the predictions of the causal responsibility attribution model, let's note that the predictions of standard preferences are straightforward: since punishment and reward is costly, the unique sequential Nash equilibrium is that the public does not react to the politicians actions and, thus, all politicians vote *no* and the reform is not implemented. This prediction is independent of the specific voting rule.

Next, we assume that the electoral public ("agent R") possesses preferences for causal responsibility attribution. In the *responsibility-attribution stage*, agent R chooses an allocation for each politician i, p_i. Due to the higher payoff, agent R prefers that the reform is implemented, $j_{R,reform}(\varnothing) > 0 > j_{R,no\,reform}(\varnothing)$. A politician who votes *yes* (*no*) has positive causal responsibility for (not) implementing the reform. Since agent R prefers the reform, he will reward and punish agents for voting *yes* and *no*, respectively, depending on their level of causal responsibility for the respective outcome. Agent R's optimal allocation for politician i after terminal history z is:

$$p_i^*(a_i, a_{-i}) = \begin{cases} \frac{\rho}{c}\, r_{i,reform}(yes, a_{-i})\; j_{R,reform}(\varnothing) & \text{if} \quad a_i = yes \\ \frac{\rho}{c}\, r_{i,no\,reform}(no, a_{-i})\; j_{R,no\,reform}(\varnothing) & \text{if} \quad a_i = no \end{cases} \qquad \forall\, i \in I$$

Throughout this section, I assume that the interval of possible allocations for i, $[\underline{p}, \overline{p}]$, is wide enough such that it yields an interior solution. And, for ease of exposition, I focus on pure strategies only. Since, in equilibrium, agent R's beliefs are correct, ex ante and overall responsibility coincide with ex post causal responsibility.

Politicians rationally anticipate the allocation strategy of the electorate and choose their strategies to maximize their expected payoff from the game as a whole. Hence, they weigh the benefits of voting *no* in the *collective-action stage* with the costs of lower allocations in the *responsibility-attribution stage*. Overall causal responsibility of a politician who voted *yes* (*no*) for (not) implementing the reform is maximal if exactly t ($t-1$) politicians vote *yes*. In this case, he is pivotal for the respective event. Accordingly, punishment and reward is maximal in these cases. Thus, if the punishment and reward is high enough to outweigh the benefits of voting *no* in this case, the allocation will be strong enough to incentivize the politician to vote *yes* in some interval around pivotality. Figure 10 shows such a situation. The line shows the number of votes for the reform, the total number of politicians n, and the voting thresholt t. The blue interval indicates the interval in which responsibility attribution is high enough to incentivize politicians to vote "yes". In the depicted case, there are two equilibria. One in which the reform is implemented (green dot), and one in which it is not implemented (red dot).

Causal responsibility attribution, thus, can sustain an equilibrium in which the reform is implemented. However, in the above example, there still exists an equilibrium in which the reform is not implemented. This equilibrium can be avoided by designing the voting institutions

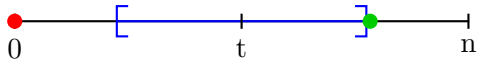


Figure 10: Equilibria with responsibility allocation and majority voting
Notes: Line shows number of "yes" votes and the implementation threshold t.

in different ways.

The first two possibilities to achieve a unique equilibrium in which the reform is implemented, simply take the same voting principle but change the parameters. First, one could simply lower the voting threshold t. In doing so, the interval in which voting *no* is discouraged moves with the threshold to the left. If t is reduced to the point in which that interval includes the case in which all politicians vote *no* (and thus responsibility and punishment for not implementing the reform is minimized), then that unwanted equilibrium ceases to exist (see Figure 11). Second, if one wants to keep the voting rule (i.e., "majority" voting) intact, one could reduce the number of politicians that get to vote on the reform in proportion with the voting threshold, e.g., by setting up a smaller committee (see Figure 12).

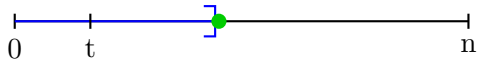


Figure 11: Equilibria with responsibility allocation and "minority" voting
Notes: Line shows number of "yes" votes and the implementation threshold t.

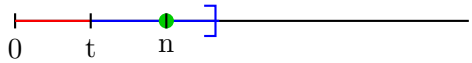


Figure 12: Equilibria with responsibility allocation and a small committee
Notes: Line shows number of "yes" votes and the implementation threshold t.

Another possibility is to change the voting rule altogether and move from a voting threshold to a consensus rule. With such a rule, the politicians have to discuss until all of them either agree on implementing the reform or on not implementing it. In this way, responsibility for both possible events is maximal since, whatever event is implemented, all politicians are pivotal for it. As we have already seen, voting *no* is discouraged in case it leads to full causal responsibility for not implementing the reform and thus, also under this change, there exists a unique equilibrium in which the reform is implemented (see Figure 13).

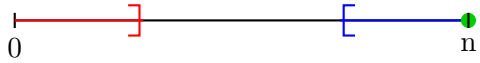


Figure 13: Equilibria with responsibility allocation and consensus rule
Notes: Line shows number of "yes" votes and the implementation threshold t.

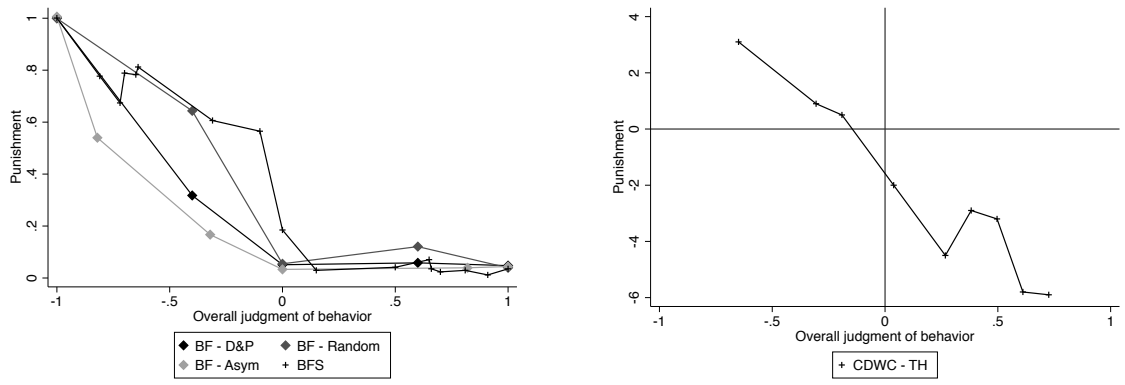
As demonstrated causal responsibility attribution can be a useful guide when designing important institutions such as voting rules. In the following, we seek to understand whether causal responsibility attribution can also help us understand behavioral phenomena that we observe in markets.

6.3 Discussion of related experimental findings

In this section, I discuss experimental findings from several studies that are suited for an application of the theory. All chosen studies consist of a first stage in which multiple agents and potentially nature implement some event and a second stage in which an agent independently decides about an allocation decision for those agents, similar to the setup of the model.

The main goal is to make comparative-statics predictions for the allocation decisions in the responsibility- attribution stage and compare those to the actual data. Therefore, I'm focussing on those treatments of the studies in which not only the extreme (e.g., being pivotal for a bad vs being pivotal for a good event), but also intermediate levels of judgment of behavior are possible. In order to get comparable predictions, I naturally have to make assumptions which I apply equally across all studies. First, I assume that agent R gives equal weight to ex ante and ex post causal responsibility ($\gamma_R = 0.5$). Second, whenever two events are possible, I assume that the reference payoff lies in the middle of the two and normalize judgment to $j_{R,x}(\cdot) = -1$ for the bad and $j_{R,x}(\cdot) = 1$ for the good event. Third, I assume that, whenever applicable, agent R's beliefs about the behavior strategies coincide with the actual frequencies with which actions are taken. For expositional ease, I assume that agent R faces a convex allocation cost function to ensure that an interior solution exists.

Figure 14 summarizes the main results. The x-axis shows the overall judgment of behavior as prescribed by the theory. A judgment of -1 or 1 means that the subject bears full causal responsibility for the bad or good event, respectively. The theory predicts punishment for negative and reward for positive judgment of behavior. Furthermore, it predicts that punishment (reward) is decreasing (increasing) the better the judgment of behavior. The y-axis depicts the actual average punishment of all subjects that bear the respective overall judgment in the respective treatment. Panel 14(a) of Figure 14 shows experiments that only allowed punishment and Panel 14(b) of Figure 14 shows data from an experiment that allowed both punishment and reward. To increase comparability, the highest observed average punishment is normalized to one and the other levels are scaled accordingly in the left panel.



(a) Punishment only.

(b) Punishment and reward.

Figure 14: Punishment and judgment of behavior in various experiments.
Notes: Ex ante causal responsibility is evaluated using the probabilities with which agents chose actions in the experiments. Punishment is normalized such that the highest average punishment is set to 1. Parameters: $\gamma_R = 0.5$, $j(good) = 1$ and $j(bad) = -1$.

A couple of interesting observations immediately stand out. First, as predicted, in all five treatments of the three experiments, punishment was always greatest when judgment of behavior was most negative. Second, as predicted, punishment was always positive for negative judgments, and close to zero (Panel 14(a)) or negative (Panel 14(b)) for positive judgments. Third, as predicted, punishment decreases for more positive judgments in all treatments. In the interest of brevity, I summarize the intuition behind the analysis of each study in the following and relegate a summary of the formal analyses to Appendix A.2.

Delegation *(Bartling and Fischbacher, 2012, BF).* In their well-known study on delegation and responsibility, Bartling and Fischbacher (2012) let dictators choose between implementing an unfair or fair payoff allocation by themselves, or delegating that decision to an intermediary. The unfair allocation gives the dictator and an intermediary a higher payoff than two recipients whereas the fair allocation equalizes all payoffs at an intermediate level. In a second stage, the recipients could attribute costly punishment points to the dictator and the intermediary after observing the choices of both players. Treatments varied whether delegation was to another subject (BF - D&P), to a lottery (BF - Random), or whether the dictator had an asymmetric choice set and could only implement the fair but not the unfair allocation himself (BF - Asym).

The authors find that in treatments "BF - D&P" and "BF - Random" the dictator is punished less when the unfair allocation is implemented after delegation compared to when he implemented it himself. Causal responsibility attribution can explain this finding. The dictator bears full causal responsibility for implementing the unfair allocation himself as he could have chosen the fair allocation instead. If he choses to delegate, he still bears full ex post causal responsibility for the allocation that the intermediary implements, as, again, he could have implemented a different allocation himself. However, his ex ante causal responsibility for the implemented allocation is reduced as it depends on the probability with which the intermediary (or the lottery) implements the respective allocation. Hence, overall causal responsibility is reduced compared to the case without delegation which leads to a better overall judgment of behavior. In addition, after delegation, the dictator bears ex ante causal responsibility for the allocation that was not implemented but had some probability of being implemented. Taken together, this predicts that the dictator's behavior is judged worse when he implements the unfair allocation himself compared to when it is implemented after delegation. The same reasoning holds for treatment "BF - Asym" with the only difference that the dictator bears no causal responsibility for the fair allocation if it is implemented after delegation, because he could not have implemented a different allocation himself.

The intermediary has full causal responsibility for whatever allocation he implements after being delegated the decision, but no causal responsibility when the decision was not delegated. Causal responsibility attribution therefore predicts that the recipient should punish a dictator who implements the unfair allocation on his own as much as an intermediary who implements the unfair allocation after delegation. Indeed, the study finds no significant difference in punishment between the two cases. Thus, for each treatment, causal responsibility attribution can explain the observed punishment pattern.

Outcome-based and intention-based social preference theories cannot explain these results to the same degree. Outcome-based social preferences (e.g. Fehr and Schmidt, 1999) would

predict punishment of the dictator and the recipient after the implementation of an unfair allocation. However, since punishment only serves to reduce payoff inequalities, the theory does not predict whether dictator or intermediary should be punished more. Intention-based social preferences (e.g. Rabin, 1993; Dufwenberg and Kirchsteiger, 2004) predict lower punishment for the dictator after delegation, but, since the evaluation of the intention does not depend on the finally implemented outcome, no difference in punishment depending on whether the delegate/lottery implemented the unfair or fair allocation. An econometric comparison of the different punishment motives reveals that causal responsibility based judgment of the agents' actions remains a highly significant predictor of punishment even after controlling for other motives (see Table A.IV in Appendix A.2).

Sequential voting *(Bartling, Fischbacher, and Schudy, 2015a, BFS).* This experiment consists of a collective action stage in which three dictators *sequentially* vote to implement either an unequal or an equal payoff allocation between them and three recipients. Each dictator observes the votes that were previously cast. The unequal allocation is implemented if at least two dictators vote for it; otherwise, the equal allocation is implemented. In a second stage, the three recipients independently assign costly punishment to each of the three dictators after observing the votes and the implemented allocation (via strategy method). Figure 15 depicts the game tree of the collective-action stage.

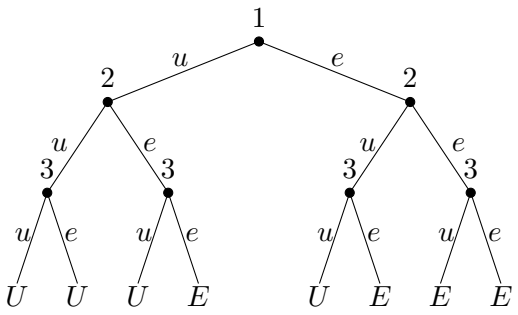


Figure 15: Collective-action stage in BFS

The experiment provides an interesting study of how the sequentiality of the voting procedure can affect causal responsibility ratings.[26] To begin with note that only the third dictator can have full causal responsibility for the unequal or equal allocation. The first is the case after histories (ueu) and (euu), the latter after histories (uee) and (eue). As predicted by the theory, average punishment is never higher than for the third dictator when he has full causal responsibility for the unequal allocation. The first and second dictator are never fully causally responsible for implementing the unequal allocation when they chose u. This is the case because both ex ante as well as ex post causal responsibility depends on the punisher's beliefs about the behavioral strategies of all dictators. I assume that these beliefs correspond to the actual frequencies with which actions were chosen. For example, after (euu), the second dictator is fully ex post causally responsible for the unequal allocation, but his ex ante causal responsibility is below 1 as the third dictator chose u only with 63 percent probability. Figure 14 shows that the theory overall

[26] Table A.VI in Appendix A.2 provides an overview over all causal responsibility ratings, the according judgment of behavior, and average punishment.

predicts the punishment pattern well.

The experiment is also a good illustration of when the predictions of the theory are less intuitive. In particular, after action profile (uee), the first dictator is still punished considerably even though the judgment of behavior is almost neutral. The overly good judgment stems from the fact that the dictator is partially ex post causally responsible for the implemented equal outcome, even though he chose u. This is the case because with some probability, the second and third dictator would have implemented the unequal outcome if the first dictator would have chosen e instead. Thus, with some probability the first dictator is pivotal for the equal outcome after (uee) even though he chose u. It is unlikely that the participants, when making their punishing choices, thought about this special counterfactual case and thus punishment is higher than predicted. Nevertheless, the corresponding punishment is still lower than for the other action profiles after the first dictator chose u and thus the comparative-statics predictions still hold up well.

For reasons similar to the ones outlined before, pure outcome- or intention-based models cannot explain the results as well. An econometric comparison of the different punishment motives reveals that causal responsibility based judgment of the agents' actions remains a highly significant predictor of punishment even after controlling for other motives (see Table A.VII in Appendix A.2).

Outcome bias *(Cushman, Dreber, Wang, and Costa, 2009, CDWC).* In this experiment, a dictator chooses one of three possible lotteries, each of which implements a different probability distribution over a *selfish*, a *fair*, and a *generous* payoff allocation between the dictator and a recipient. Each of the three lotteries assigns 2/3 probability to one of the allocations and 1/6 probability to the other two allocations. Hence, if the dictator wants to maximize the probability of implementing the *selfish* allocation, he chooses the lottery that selects the *selfish* allocation with 2/3 probability. After observing the choice of the dictator and the outcome of the lottery, the recipient can punish and reward the dictator at no cost to himself.

The results show that the dictator is punished for a *selfish* and rewarded for a *fair* or *generous* allocation whenever they occur. However, the amount of punishment and reward is moderated by the choice of the lottery. Punishment is larger and reward is lower when the dictator chose the lottery with the highest probability of yielding the *selfish* allocation.

Hence, outcome- or intention-based social preferences alone cannot explain the result. Causal responsibility can rationalize the behavior of the recipient in the following way: Ex ante causal responsibility depends solely on the choice of the lottery and not on the final allocation. Ex ante causal responsibility is always highest for the allocation that the chosen lottery selects with the highest probability. This explains the finding that the same allocation leads to different reward and punishment depending on the chosen lottery. Ex post causal responsibility is only positive for the allocation that is implemented by the lottery. Hence, this explains that for a given choice of lottery the implemented allocation still influences punishment and reward.

Inequality acceptance & redistribution *(Cappelen, Fest, Sørensen, and Tungodden, 2020).* Preferences for causal responsibility attribution can also influence inequality acceptance and redistributive choices. For example, when inequality arises due to people's actual choices, they are causally responsible for the resulting inequality and, thus, other people might be less willing

to engage in ex post redistribution to decrease the inequality. On the other hand, if no choice is possible and inequality arises due to luck (e.g., inheritances), they are not causally responsible for the inequality and thus other people might be more willing to engage in redistribution.

Cappelen et al. (2020) study this question with a large-scale sample of the general population of Norway. The study has a baseline and two choice treatments. In the baseline, inequality is completely determined by luck: a lottery gives all the earnings to one of two stakeholders and no earnings to the other. In the two choice treatments, the stakeholders make a choice before their earnings are determined. In the "nominal choice" treatment, they choose between two lotteries. The lotteries are identical ex ante and, thus, stakeholders cannot change the likelihoods of the possible outcomes with their choice. In the "forced choice" treatment, stakeholders choose between a lottery that is identical to the lottery in the baseline and a safe alternative. The safe alternative, however, is no acceptable alternative to the lottery as it is close to zero. Thus, in both treatments, participants do choose, but their choice is almost meaningless.

In all treatments, impartial spectators can then redistribute the earnings between the two stakeholders. The authors find that impartial spectators accept a significantly higher level of inequality, if a stakeholder had a choice compared to the baseline. Furthermore, inequality acceptance was significantly larger in the "forced choice" compared to the "nominal choice" treatment. Using additional survey data, the authors rule out potential explanations such as illusion of control, fundamental attribution error, misunderstanding and intuitive decision making.

However, this counterintuitive finding can be explained by ex post causal responsibility if one assumes that spectators have a preference to reduce inequality when the stakeholders have no causal responsibility for it, but not when the stakeholders are causally responsible for it. In the baseline, both stakeholders have no choice to make and can thus never be pivotal for the outcome. Thus, their causal responsibility for the resulting inequality is zero and spectators feel free to redistribute. In the "nominal choice" treatment, a stakeholder chooses one of two identical lotteries. If she is lucky or unlucky, she has a partial ex post causal responsibility for her high/low payoff as she would have, in expectation, a lower/higher payoff had she chosen the other lottery. Thus, if a spectator places some weight on ex post causal responsibility, a reduced redistribution compared to the baseline is justified. In the "forced choice" treatment, a lucky stakeholder has full ex post causal responsibility for the outcome, as he would have been worse off with certainty, had he chosen the safe alternative. An unlucky stakeholder, on the other hand, has no causal responsibility because he could not have improved by choosing the safe alternative. Thus, compared to the "nominal choice" treatment, responsibility for the lucky outcome is higher in the forced choice treatment. This can explain the significantly lower redistribution in this treatment compared to both other treatments.[27]

7 Conclusion

This paper introduces and provides evidence for a versatile notion of causal responsibility that can be applied to simultaneous- and sequential-move games. An agent with responsibility pref-

[27] Table A.VIII in Appendix A.2 summarizes the causal responsibility levels for different outcomes of the lottery.

erences cares about his own or others' causal responsibility for events into account when deciding about his actions. Applications demonstrate that taking responsibility preferences into account can matter for worker's provision of effort and the design of important institutions such as voting rules. Finally, the paper tests the predictive power of the causal responsibility notion for allocation decisions in data from existing laboratory experiments; it finds that the causal responsibility motive can explain observed punishment patterns successfully.

The paper represents a starting point for the study of causal responsibility attributions in economics. To conclude, I will therefore discuss several promising directions for future theoretical and empirical research. For example, the perception of causality might be subject to biases. Kahneman and Miller (1986) propose that the causal impact of behavior on an event is more salient to people when the behavior deviates from what is expected as normal behavior. Following Ross (1977), a literature on the *fundamental attribution error* suggests that people judge internal factors of other people (e.g., their character) as more causal for events than external factors. Weber et al. (2001) have demonstrated experimentally this effect by showing that leaders are wrongly perceived as more causal for a group's coordination outcome compared to external factors (in this case the group size). A structured empirical analysis of which factors might bias those perceptions would be worthwhile.

Furthermore, causal responsibility is only one among several predictors of punishment and reward. While the point of this paper was to demonstrate the usefulness of a causal responsibility concept, future work could integrate responsibility with other motives, such as intentions, into a single theoretical framework. Empirically it will be important to understand how these different predictors interact and in which situations one predictor works better than another. In addition, aside from causal reasoning, several other factors can affect responsibility perceptions and the attribution of blame and praise. For example, people are prone to react more strongly to negative events that are implemented as a main effect compared to when they occur as side effects (Knobe, 2003). Furthermore, increased spatial, temporal, and social distance might reduce perceptions of responsibility (for a discussion of these channels, see Greene (2013)). For example, Coffman (2011) and Oexl and Grossman (2012) show that delegation reduces punishment even when the intermediary has no choice and thus causal responsibility is not reduced. They explain this with the increased distance between the actor and the victim. Additionally, people perceive the implementation of a bad event as less bad when it was implemented due to an omission (not changing the status quo), as opposed to when it was implemented due to a commission (changing the status quo) (Cox et al., 2016). Bartling et al. (2014) show that willful ignorance can reduce punishment even when it doesn't have a causal impact. Bartling and Özdemir (2017) provide evidence that deviations from pivotality are especially relevant if there does not exist a strong social norm for the "morally correct" course of action. All of these points raise interesting questions for future research.

References

Alicke, M. D., D. R. Mandel, D. J. Hilton, T. Gerstenberg, and D. A. Lagnado (2015): "Causal Conceptions in Social Explanation and Moral Evaluation: A Historical Tour," *Perspectives on Psychological Science*, 10, 790–812.

Anselm, R., D. Bhatia, U. Fischbacher, and J. Hausfeld (2022): "Blame and Praise: Responsibility Attribution Patterns in Decision Chains," *Thurgau Institute of Economics Research Paper Series*, 126.

Bartling, B., F. Engl, and R. A. Weber (2014): "Does willful ignorance deflect punishment? – An experimental study," *European Economic Review*, 70, 512–524.

Bartling, B. and U. Fischbacher (2012): "Shifting the blame: on delegation and responsibility," *Review of Economic Studies*, 79, 67–87.

Bartling, B., U. Fischbacher, and S. Schudy (2015a): "Pivotality and Responsibility Attribution in Sequential Voting," *Journal of Public Economics*, 128, 133–139.

Bartling, B. and Y. Özdemir (2017): "The Limits to Moral Erosion in Markets: Social Norms and the Replacement Excuse," *Working Paper*.

Bartling, B., R. A. Weber, and L. Yao (2015b): "Do Markets Erode Social Responsibility?" *The Quarterly Journal of Economics*, 130, 219–266.

Battigalli, P. and M. Dufwenberg (2007): "Guilt in Games," *American Economic Review*, 97, 170–176.

——— (2009): "Dynamic psychological games," *Journal of Economic Theory*, 144, 1–35.

Battigalli, P., M. Dufwenberg, and A. Smith (2016): "Frustration & Anger in Games," *Working Paper*, 1–44.

Beebee, H., C. Hitchcock, and P. Menzies, eds. (2012): *The Oxford Handbook of Causation*, Oxford University Press.

Behnk, S., L. Hao, and E. Reuben (2017): "Partners in Crime: Diffusion of Responsibility in Antisocial Behaviors," *IZA Discussion Paper*, 11031.

——— (2019): "Shifting normative beliefs: On why groups behave more antisocially than individuals," *Working Paper*.

Bénabou, R., A. Falk, and J. Tirole (2018): "Narratives, imperatives, and moral reasoning," Tech. rep., National Bureau of Economic Research.

Bénabou, R. and J. Tirole (2003): "Intrinsic and Extrinsic Motivation," *The Review of Economic Studies*, 70, 489–520.

Bolton, G. E. and A. Ockenfels (2000): "ERC: A Theory of Equity, Reciprocity, and Competition," *American Economic Review*, 90, 166–193.

Brütt, K., A. Schram, and J. Sonnemans (2020): "Endogenous group formation and responsibility diffusion: An experimental study," *Games and Economic Behavior*, 121, 1–31.

Cappelen, A. W., S. Fest, E. Ø. Sørensen, and B. Tungodden (2020): "Choice and Personal Responsibility: What Is a Morally Relevant Choice?" *The Review of Economics and Statistics*, 1–35.

Cappelen, A. W., J. Konow, E. Ø. Sørensen, and B. Tungodden (2013): "Just Luck: An Experimental Study of Risk-Taking and Fairness," *American Economic Review*, 103, 1398–1413.

Cappelen, A. W., E. Ø. Sørensen, and B. Tungodden (2010): "Responsibility for what? Fairness and individual responsibility," *European Economic Review*, 54, 429–441.

Cappelen, A. W. and B. Tungodden (2006): "Relocating the responsibility cut: Should more responsibility imply less redistribution?" *Politics, Philosophy & Economics*, 5, 353–362.

Cason, T. N. and V.-L. Mui (1997): "A Laboratory Study of Group Polarisation in the Team Dictator Game," *The Economic Journal*, 107, 1465–1483.

Cassar, L. and S. Meier (2018): "Nonmonetary incentives and the implications of work as a source of meaning," *Journal of Economic Perspectives*, 32, 215–38.

Charness, G. (2000): "Responsibility and effort in an experimental labor market," *Journal of Economic Behavior & Organization*, 42, 375–384.

——— (2004): "Attribution and Reciprocity in an Experimental Labor Market," *Journal of Labor Economics*, 22, 665–688.

Chockler, H. and J. Halpern (2004): "Responsibility and Blame: A Structural-Model Approach," *J. Artif. Intell. Res.(JAIR)*, 22, 93–115.

Choo, L., V. Grimm, G. Horváth, and K. Nitta (2019): "Whistleblowing and diffusion of responsibility: An experiment," *European Economic Review*, 119, 287–301.

Coase, R. H. (1960): "The Problem of Social Cost," *The Journal of Law and Economics*, 3, 1–44.

Coffman, L. C. (2011): "Intermediation Reduces Punishment (and Reward)," *American Economic Journal: Microeconomics*, 3, 77–106.

Cox, J. C., M. Servátka, and R. Vadovič (2016): "Status quo effects in fairness games: reciprocal responses to acts of commission versus acts of omission," *Experimental Economics*, 1–18.

Cryder, C. E. and G. Loewenstein (2012): "Responsibility: The tie that binds," *Journal of Experimental Social Psychology*, 48, 441–445.

Cushman, F. (2008): "Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment." *Cognition*, 108, 353–80.

Cushman, F., A. Dreber, Y. Wang, and J. Costa (2009): "Accidental outcomes guide punishment in a "trembling hand" game." *PloS one*, 4, e6699.

Dana, J., R. A. Weber, and J. X. Kuang (2007): "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness," *Economic Theory*, 33, 67–80.

Darley, J My, J. and B. Latané (1968): "Bystander Intervention in Emergencies: Diffusion of Responsibility," *Journal of Personality and Social Psychology*, 8, 377–383.

Darley, J. M. and T. R. Shultz (1990): "Moral Rules: Their Content and Acquisition," *Annual Review of Psychology*, 41, 525–556.

Duch, R., W. Przepiorka, and R. Stevenson (2014): "Responsibility attribution for collective decision makers," *American Journal of Political Science*, 59, 372–389.

Dufwenberg, M. and G. Kirchsteiger (2004): "A theory of sequential reciprocity," *Games and Economic Behavior*, 47, 268–298.

Ellingsen, T. and M. Johannesson (2008): "Pride and prejudice: The human side of incentive theory," *American economic review*, 98, 990–1008.

Epstein, R. A. (1973): "A Theory of Strict Liability," *Journal of Legal Studies*, 151–204.

Falk, A., T. Neuber, and N. Szech (2020): "Diffusion of Being Pivotal and Immoral Outcomes," *The Review of Economic Studies*, 87, 2205–2229.

Fehr, E. and K. M. Schmidt (1999): "A Theory of Fairness, Competition, and Cooperation," *The Quarterly Journal of Economics*, 114, 817–868.

Gawn, G. and R. Innes (2021): "Machiavelli Preferences Without Blame: Delegating Selfish vs. Generous Decisions in Dictator Games: Machiavelli Preferences Without Blame," *Journal of Behavioral and Experimental Economics*, 90.

Gerstenberg, T., N. D. Goodman, D. a. Lagnado, and J. B. Tenenbaum (2014): "From Counterfactual Simulation to Causal Judgment," *Proceedings of the 36th Annual Conference of the Cognitive Science Society (CogSci 2014)*, 1, 523–528.

Gerstenberg, T. and D. a. Lagnado (2010): "Spreading the blame: The allocation of responsibility amongst multiple agents." *Cognition*, 115, 166–71.

——— (2012): "When contributions make a difference: explaining order effects in responsibility attribution." *Psychonomic bulletin & review*, 19, 729–36.

Gerstenberg, T., M. F. Peterson, N. D. Goodman, D. A. Lagnado, and J. B. Tenenbaum (2017): "Eye-Tracking Causality," *Psychological Science*, 1–14.

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

Green, S. (2015): *Causation in Negligence*, Oxford: Hart Publishing.

Greene, J. (2013): *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*, New York: Penguin Press.

Guglielmo, S. (2015): "Moral judgment as information processing: An integrative review," *Frontiers in Psychology*, 6, 1–19.

Halpern, J. Y. and J. Pearl (2005): "Causes and Explanations: A Structural-Model Approach. Part I: Causes," *The British Journal for the Philosophy of Science*, 56, 843–887.

Hamman, J. R., G. Loewenstein, and R. A. Weber (2010): "Self-Interest through Delegation: An Additional Rationale for the Principal-Agent Relationship," *American Economic Review*, 100, 1826–1846.

Hart, H. L. A. (1968): *Punishment and Responsibility*, Oxford University Press, 1st ed.

Hart, H. L. A. and A. Honoré (1959): *Causation in the Law*, Oxford: Clarendon Press, 1st ed.

Heider, F. (1958): *The Psychology of Interpersonal Relations*, New York, NY: Wiley.

Hume, D. (1777): *An Enquiry Concerning Human Understanding*, London: Millar, A.

Kahneman, D. and S. Frederick (2002): "Representativeness Revisited: Attribute Substitution in Intuitive Judgment," in *Heuristics and Biases: The Psychology of Intuitive Judgment*, ed. by T. Gilovich, D. W. Griffin, and D. Kahneman, New York: Cambridge University Press, 49–81.

Kahneman, D. and D. T. Miller (1986): "Norm theory: Comparing reality to its alternatives." *Psychological Review*, 93, 136–153.

Kahneman, D. and A. Tversky (1982): "The simulation heuristic," in *Judgment under uncertainty: Heuristics and biases*, ed. by D. Kahneman, P. Slovic, and T. A., New York: Cambridge University Press, 201–208.

Kahneman, D. and C. a. Varey (1990): "Propensities and counterfactuals: The loser that almost won." *Journal of Personality and Social Psychology*, 59, 1101–1110.

Kelley, H. H. (1967): "Attribution theory in social psychology," in *Nebraska Symposium on Motivation*, ed. by D. Levine, Lincoln: University of Nebraska Press, vol. 15, 192–240.

Kirchkamp, O. and C. Strobel (2019): "Sharing responsibility with a machine," *Journal of Behavioral and Experimental Economics*, 80, 25–33.

Kirchler, M., J. Huber, M. Stefan, and M. Sutter (2016): "Market Design and Moral Behavior," *Management Science*, 62, 2457–2764.

Knobe, J. (2003): "Intentional action in folk psychology: An experimental investigation," *Philosophical Psychology*, 16, 309–324.

Kocher, M. G., S. Schudy, and L. Spantig (2018): "I Lie? We Lie! Why? Experimental Evidence on a Dishonesty Shift in Groups," *Management Science*, 64, 3995–4008.

Lagnado, D. a., T. Gerstenberg, and R. Zultan (2013): "Causal responsibility and counterfactuals," *Cognitive science*, 37, 1036–73.

Landes, W. and R. Posner (1983): "Causation in Tort Law: An Economic Approach," *The Journal of Legal Studies*, 12, 109–134.

Leslie, A. M. and S. Keeble (1987): "Do six-month-old infants perceive causality?" *Cognition*, 25, 265–288.

Lewis, D. (1973): "Causation," *The Journal of Philosophy*, 70, 556–567.

——— (1986): *Philosophical Papers: Volume II*, Oxford University Press.

Luhan, W. J., M. G. Kocher, and M. Sutter (2009): "Group polarization in the team dictator game reconsidered," *Experimental Economics*, 12, 26–41.

Mackie, J. L. (1974): *The Cement of the Universe: A Study of Causation*, Oxford, United Kingdom: Oxford University Press.

Malle, B. F., S. Guglielmo, and A. E. Monroe (2014): "A Theory of Blame," *Psychological Inquiry*, 25, 147–186.

Mandel, D. R. and D. R. Lehman (1996): "Counterfactual Thinking and Ascriptions of Cause and Preventability." *Journal of Personality and Social Psychology*, 71, 450–463.

Manove, M. (1997): "Job Responsibility, Pay and Promotion," *The Economic Journal*, 107, 85–103.

McGill, a. L. and a. E. Tenbrunsel (2000): "Mutability and propensity in causal selection." *Journal of Personality and Social Psychology*, 79, 677–89.

Michotte, A. (1963): *The perception of causality*, England: Oxford University Press.

Moore, M. S. (2009): *Causation and responsibility: An essay in law, morals, and metaphysics*, Oxford University Press.

Oexl, R. and Z. J. Grossman (2012): "Shifting the blame to a powerless intermediary," *Experimental Economics*, 16, 306–312.

Pearl, J. (2000): *Causality: Models, Reasoning and Inference*, Cambridge University Press.

Prendergast, C. J. (1995): "A Theory of Responsibility in Organizations," *Journal of Labor Economics*, 13, 387–400.

Rabin, M. (1993): "Incorporating Fairness into Game Theory and Economics," *American Economic Review*, 83, 1281–1302.

Roese, N. J. (1997): "Counterfactual Thinking," *Psychological Bulletin*, 121, 133–48.

Ross, L. (1977): "The Intuitive Psychologist And His Shortcomings: Distortions in the Attribution Process," .

Rothenhäusler, D., N. Schweizer, and N. Szech (2018): "Guilt in voting and public good games," *European Economic Review*, 101, 664–681.

Saito, K. (2013): "Social Preferences under Risk: Equality of Opportunity versus Equality of Outcome," *American Economic Review*, 103, 3084–3101.

Saxe, R. and S. Carey (2006): "The perception of causality in infancy," *Acta Psychologica*, 123, 144–165.

Schelling, T. C. (1980): *The Strategy of Conflict*, Cambridge, MA: Harvard University Press.

Shaver, K. G. (1985): *The Attribution of Blame: Causality, Responsibility, and Blameworthiness*, Springer.

Shultz, T. R., M. Schleifer, and I. Altman (1981): "Judgments of Causation, Responsibility, and Punishment in Cases of Harm-Doing," *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, 13, 238–253.

Sliwka, D. (2006): "On the notion of responsibility in organizations," *Journal of Law, Economics, and Organization*, 22, 523–547.

——— (2007): "Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes," *American Economic Review*, 97, 999–1012.

Smith, A. (1759): *The Theory of Moral Sentiments*, London, Edinburgh: A. Miller, A. Kincaid and J. Bell.

Spellman, B. (1997): "Crediting Causality," *Journal of Experimental Psychology: General*, 126, 323–348.

Steel, S. (2015): *Proof of Causation in Tort Law*, Cambridge University Press.

Summerville, A. and N. J. Roese (2008): "Dare to compare: Fact-based versus simulation-based comparison in daily life," *Journal of Experimental Social Psychology*, 44, 664–671.

Tungodden, B. (2005): "Responsibility and redistribution: The case of first best taxation," *Social Choice and Welfare*, 24, 33–44.

Tversky, A. and D. Kahneman (1982): "Causal schemas in judgments under uncertainty," in *Judgment under uncertainty: Heuristics and biases*, ed. by D. Kahneman, P. Slovic, and T. A., New York: Cambridge University Press, 117–128.

Von Siemens, F. A. (2013): "Intention-based reciprocity and the hidden costs of control," *Journal of Economic Behavior & Organization*, 92, 55–65.

Weber, R., C. Camerer, Y. Rottenstreich, and M. Knez (2001): "The illusion of leadership: Misattribution of cause in coordination games," *Organization Science*, 12, 582–598.

Weiner, B. (1995): *Judgments of Responsibility: A Foundation for a Theory of Social Conduct*, The Guilford Press.

Wells, G. L. and I. Gavanski (1989): "Mental simulation of causality," *Journal of Personality and Social Psychology*, 56, 161–169.

Woodward, J. (2003): *Making Things Happen: A Theory of Causal Explanation*, Oxford University Press.

Wright, R. W. (1985): "Causation in Tort Law," *California Law Review*, 73, 1735–1828.

——— (1988): "Causation, Responsibility, Risk, Probability, Naked Statistics, and Proof: Pruning the Bramble Bush by Clarifying the Concepts," *Iowa Law Review*, 73, 1002–1077.

Zultan, R., T. Gerstenberg, and D. a. Lagnado (2012): "Finding fault: causality and counterfactuals in group attributions," *Cognition*, 125, 429–40.

A Appendix

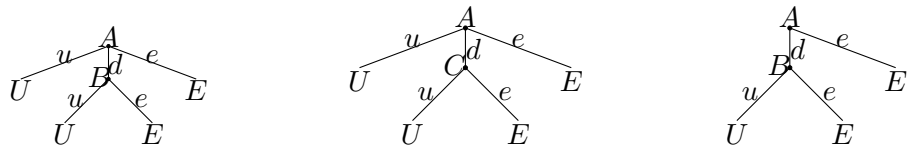
A.1 Experimental evidence

Table A.I: Summary of sessions.

	Session 1	Session 2	Session 3	Session 4
Date	10/19/18	10/19/18	11/02/18	11/02/18
Time	10-10.45	11.45-12.20	10-10.45	11.15-12
Participants	30	30	13	26
Scenario 1	yes	yes	yes	yes
Scenario 2	yes	yes	no	no
Scenario 3	no	no	yes	yes
Scenario 4	no	no	yes	yes

A.2 Related experiments

Delegation *(Bartling and Fischbacher, 2012, BF)*
Figure 16 shows the three analyzed treatments:



(a) D&P (b) Random (c) Asymmetric

Figure 16: Treatments with variation in causal responsibility.

- D&P
  - Stage-1 agents: $I = \{A, B\}$
  - Action sets: $A_A(\varnothing) = \{u, e, d\}$, $A_B(d) = \{u, e\}$
  - Beliefs: $\alpha_{KB} = \sigma_B = (0.4, 0.6)$
- Random
  - Stage-1 agents: $I = \{A\}$, $I_C = \{A, C\}$
  - Action sets: $A_A(\varnothing) = \{u, e, d\}$, $A_C(d) = \{u, e\}$
  - Beliefs: $\alpha_{KC} = \sigma_C = (0.4, 0.6)$
- Asym
  - Stage-1 agents: $I = \{A, B\}$
  - Action sets: $A_A(\varnothing) = \{e, d\}$, $A_B(d) = \{u, e\}$
  - Beliefs: $\alpha_{KB} = \sigma_B = (0.64, 0.36)$
- Events and judgments: $X = \{U, E\}$, $j_U(\varnothing) = j_U(d) = -1$, $j_E(\varnothing) = j_E(d) = 1$

Table A.II: Causal responsibility ratings for player A in BF.

Treat	Pl. A	Pl. B	$r^{EP}_{A,U}$	$r^{EA}_{A,U}$	$r_{A,U}$	$r^{EP}_{A,F}$	$r^{EA}_{A,F}$	$r_{A,F}$	$\sum_x r_{A,x}() \cdot j_x()$
D&P	u	-	1	1	1	0	0	0	-1
D&P	delegate	u	1	0.4	0.7	0	0.6	0.3	-0.4
D&P	delegate	f	0	0.4	0.2	1	0.6	0.8	0.6
D&P	f	-	0	0	0	1	1	1	1
Random	u	-	1	1	1	0	0	0	-1
Random	delegate	u	1	0.4	0.70	0	0.6	0.30	-0.4
Random	delegate	f	0	0.4	0.20	1	0.6	0.80	0.6
Random	f	-	0	0	0	1	1	1	1
Asym	delegate	u	1	0.64	0.82	0	0	0	-0.82
Asym	delegate	f	0	0.64	0.32	0	0	0	-0.32
Asym	f	-	0	0	0	0.82	0.82	0.82	0.82

Table A.III: Causal responsibility ratings for player B in BF.

Treat	Pl. A	Pl. B	$r^{EP}_{B,U}$	$r^{EA}_{B,U}$	$r_{B,U}$	$r^{EP}_{B,F}$	$r^{EA}_{B,F}$	$r_{B,F}$	$\sum_x r_{B,x}() \cdot j_x()$
D&P	u	-	0	0	0	0	0	0	0
D&P	delegate	u	1	1	1	0	0	0	-1
D&P	delegate	f	0	0	0	1	1	1	1
D&P	f	-	0	0	0	0	0	0	0
Random	u	-	0	0	0	0	0	0	0
Random	delegate	u	0	0	0	0	0	0	0
Random	delegate	f	0	0	0	0	0	0	0
Random	f	-	0	0	0	0	0	0	0
Asym	delegate	u	1	1	1	0	0	0	-1
Asym	delegate	f	0	0	0	1	1	1	1
Asym	f	-	0	0	0	0	0	0	0

Table A.IV: Robustness of causal responsibility as punishment motive

Punishment	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) OLS
Causal	-3.344***	-2.983***	-4.483***	-3.090***	-6.266***
responsibility	(0.172)	(0.190)	(0.287)	(0.458)	(0.844)
Outcome		0.445***			0.321***
unfair		(0.094)			(0.095)
Intention			-1.157***		-1.364***
"unkind"			(0.199)		(0.277)
Outcome unfair X				0.249	-1.806***
Intention "unkind"				(0.424)	(0.590)
Constant	0.289***	0.181***	0.363***	0.292***	0.276***
	(0.043)	(0.044)	(0.041)	(0.043)	(0.041)
Observations	1788	1788	1788	1788	1788
Adjusted R^2	0.378	0.384	0.385	0.378	0.390

Notes: The dependent variable is attributed punishment points for the dictator or the intermediary. Besides the "Overall judgment" variable, the other explanatory variables are constructed as in Bartling and Fischbacher (2012). Robust standard errors (clustered on 274 individuals) in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

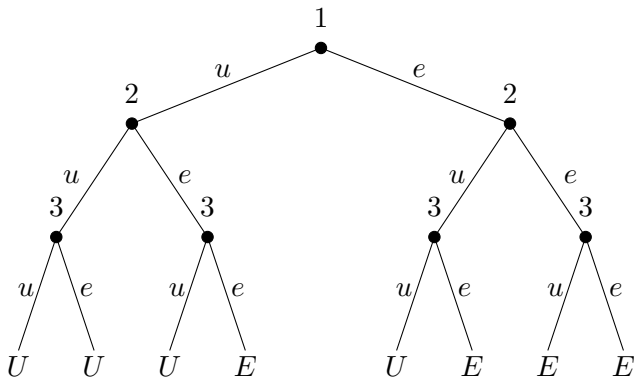
Outcome bias *(Cushman, Dreber, Wang, and Costa, 2009, CDWC).*

- Collective-action stage
  - Agents: $I = \{D\}$, $I_C = \{D, 0\}$
  - Action sets: $A_D(\varnothing) = \{a_s^D, a_f^D, a_g^D\}$, $A_0(a_s^D) = A_0(a_f^D) = A_0(a_g^D) = \{a_s^0, a_f^0, a_g^0\}$
  - Strategy of nature: $\sigma_0(a_s^D) = (\frac{2}{3}, \frac{1}{6}, \frac{1}{6})$, $\sigma_0(a_f^D) = (\frac{1}{6}, \frac{2}{3}, \frac{1}{6})$, $\sigma_0(a_g^D) = (\frac{1}{6}, \frac{1}{6}, \frac{2}{3})$
  - Events: $X = \{S, F, G\} = \{(10, 0), (5, 5), (0, 10)\}$
  - Implementation of events: $f(a_s^0) = S$, $f(a_f^0) = F$, $f(a_g^0) = G$
- Responsibility-attribution stage
  - Judgment of events: $j_S(\varnothing) = -1$ $j_F(d) = 0.5$, $j_G(\varnothing) = j_E(d) = 1$. Judgments proportionally corresponds to the punishment and reward in the "full control" condition. In that conditions, an agent was always fully causally responsible for the event and thus punishment and reward is informative about how the event was judged.
  - Beliefs: Accurate beliefs about lottery.
  - Allocation: Costfree punishment or reward (strategy method)

Table A.V: Causal responsibility ratings in CDWC.

History	$r_{D,S}^{EP}$	$r_{D,S}^{EA}$	$r_{D,S}$	$r_{D,F}^{EP}$	$r_{D,F}^{EA}$	$r_{D,F}$	$r_{D,G}^{EP}$	$r_{D,G}^{EA}$	$r_{D,G}$	$\sum_x r_{D,x}() \cdot j_x()$
(a_s^D, a_s^0)	.92	.61	.75	0	.15	.08	0	.15	.08	-0.65
(a_s^D, a_f^0)	0	.61	.31	.92	.15	.53	0	.15	.08	0.04
(a_s^D, a_g^0)	0	.61	.31	0	.15	.08	.92	.15	.53	0.27
(a_f^D, a_s^0)	.92	015	.53	0	.61	.31	0	.15	.08	-0.31
(a_f^D, a_f^0)	0	.15	.08	.92	.61	.76	0	.15	.08	0.38
(a_f^D, a_g^0)	0	.15	.08	0	.61	.31	.92	.15	.53	0.61
(a_g^D, a_s^0)	.92	.15	.53	0	.15	.08	0	.61	.31	-0.19
(a_g^D, a_f^0)	0	.15	.08	.92	.15	.53	0	.61	.31	0.50
(a_g^D, a_g^0)	0	.15	.08	0	.15	.08	.92	.61	.76	0.73

Sequential voting *(Bartling, Fischbacher, and Schudy, 2015a, BFS).*



- Collective-action stage
  - Agents: $I = \{1, 2, 3\}$
  - Action sets: $A_i(h) = \{u, e\}$ $\forall$ $i \in I$ and $h \in H$

- Events: $X = \{U, E\} = \{(9,9,9,1,1,1), (5,5,5,5,5,5)\}$
- Implementation of events: $f(\mathbf{a}) = U$ if $\sum_{i \in I} \mathbb{1}(a_i = u) \geq 2$, $f(\mathbf{a}) = F$ if otherwise

• Responsibility-attribution stage

- Judgment of events: $j_U(z) > 0 > j_E(z)$
- Beliefs: Equal to actual choice probabilities (see table).
- Allocation: Costly punishment (strategy method)

Table A.VI: Causal responsibility ratings in BFS.

History	(uuu)	(uue)	(ueu)	(uee)	(euu)	(eue)	(eeu)	(eee)
$r^{EA}_{1,U}$	.71	.71	.71	.71	.19	.19	.19	.19
$r^{EP}_{1,U}$	.80	.77	.81	.00	.39	.00	.00	.00
$r_{1,U}$	.76	.74	.76	.36	.29	.09	.09	.09
$r^{EA}_{1,E}$	.08	.08	.08	.08	.58	.58	.58	.58
$r^{EP}_{1,E}$	.00	.00	.00	.44	.00	.90	1.00	.92
$r_{1,E}$	.04	.04	.04	.26	.29	.74	.79	.75
$\sum_x r_{1,x}() \cdot j_x()$	-.72	-.70	-.72	-.10	.00	.65	.70	.66
$p_1(h)$	1.5	1.86	1.68	1.33	0.11	0.17	0.06	0.08
$r^{EA}_{2,U}$	.65	.65	.00	.00	.63	.63	.00	.00
$r^{EP}_{2,U}$	.66	.64	.00	.00	1.00	.00	.00	.00
$r_{2,U}$	.65	.64	.00	.00	.81	.31	.00	.00
$r^{EA}_{2,E}$	.00	.00	.29	.00	.00	.00	.81	.81
$r^{EP}_{2,E}$	.00	.00	.00	1.00	.00	.00	1.00	.81
$r_{2,E}$	.00	.00	.15	.50	.00	.00	.91	.81
$\sum_x r_{2,x}() \cdot j_x()$	-.65	-.64	.15	.50	-.81	-.31	.91	.81
$p_2(h)$	1.85	1.92	0.07	0.1	1.83	1.43	0.03	0.07
$r^{EA}_{3,U}$	.00	.00	1.00	.00	1.00	.00	.00	.00
$r^{EP}_{3,U}$	.00	.00	1.00	.00	1.00	.00	.00	.00
$r_{3,U}$	.00	.00	1.00	.00	1.00	.00	.00	.00
$r^{EA}_{3,E}$	.00	.00	.00	1.00	.00	1.00	.00	.00
$r^{EP}_{3,E}$	.00	.00	.00	1.00	.00	1.00	.00	.00
$r_{3,E}$	.00	.00	.00	1.00	.00	1.00	.00	.00
$\sum_x r_{3,x}() \cdot j_x()$	.00	.00	-1.00	1.00	-1.00	1.00	.00	.00
$p_3(h)$	0.86	0.26	2.39	0.08	2.33	0.08	0.92	0.03

Table A.VII: Robustness of causal responsibility as punishment motive

Punishment	(1) OLS	(2) OLS	(3) OLS	(4 OLS)	(5) OLS	(6) OLS	(7) OLS
Causal	1.956***	1.903***	1.375***	1.215***	1.606***	1.737***	0.657***
responsibility	(0.193)	(0.202)	(0.305)	(0.194)	(0.248)	(0.192)	(0.208)
Outcome		0.073					0.048
unequal		(0.113)					(0.070)
Choice unequal			0.532**				0.453***
			(0.231)				(0.161)
"Intention unkind"				0.719***			0.517**
				(0.157)			(0.197)
Choice unequal X					0.360		0.042
"Intention unkind"					(0.216)		(0.227)
"Pivotality"						0.452***	0.403**
						(0.161)	(0.155)
Constant	0.143***	0.127**	0.095***	0.122***	0.154***	0.150***	0.083**
	(0.041)	(0.057)	(0.033)	(0.041)	(0.038)	(0.041)	(0.037)
Observations	1728	1728	1728	1728	1728	1728	1728
Adjusted R^2	0.262	0.262	0.267	0.274	0.265	0.269	0.281

Notes: The dependent variable is attributed punishment points for voters. Besides the causal responsibility variable, the other explanatory variables are constructed as in Bartling et al. (2015a): *Outcome unequal* is a dummy variable which equals 1 if the unequal allocation is implemented. *Choice unequal* is a dummy variable which equals 1 if the $a_i = u$ is chosen. *"Intention unkind"* is a dummy variable equal to 1 if the respective voter opted for the unequal allocation and no majority was achieved before her vote. *"Pivotality"* is a dummy equal to 1 if the $a_i = u$ is chosen, the unequal allocation occurred, and the respective voter was the second voter opting for the unequal allocation. Robust standard errors (clustered on 72 individuals) in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.VIII: Causal responsibility ratings for a stakeholder in Cappelen et al. (2020).

Treatment	Stakeholder	Lottery	$r^{EP}_{S,hi}$	$r^{EA}_{S,hi}$	$r_{S,hi}$	$r^{EP}_{S,lo}$	$r^{EA}_{S,lo}$	$r_{S,lo}$
Baseline	-	hi	0	0	0	0	0	0
Baseline	-	lo	0	0	0	0	0	0
Nominal	$L1$	hi	0.5	0.25	0.375	0	0.25	0.125
Nominal	$L1$	lo	0	0.25	0.125	0.5	0.25	0.375
Forced	$L1$	hi	1	0.5	0.75	0	0	0
Forced	$L1$	lo	0	0.5	0.25	0	0	0