Perceived Responsibility for Negative Externalities

Horia Guias[†]

November 10, 2026

[Click Here for Most Recent Version]

Abstract

Why do some individuals pay to offset their own emissions but not the same emissions caused by someone else? I study how previously caused harm and anticipated outcomes shape the perception of responsibility for negative externalities, and how that perception affects decision-making. In a theoretical framework, a decision maker can repair personally caused harm and exogenous background harm of unknown origin. Beyond outcome-based preferences, she cares about maintaining a moral self-image that is activated by backward-looking responsibility. Under certain conditions, a desire to restore that moral self-image overrides efficiency or equity concerns, leading the decision maker to repair exclusively personally caused harm while ignoring background harm. Based on this behavioural prediction, I propose a sharp revealed preference test that isolates responsibility-based preferences from outcome-based preferences. This informs the design of an online experiment in which I vary the causal history of harm, the price to repair personally caused harm, and the price to repair background harm. Structural estimation of preferences reveals considerable heterogeneity in the influence of perceived responsibility on behaviour both at the individual and aggregate level. Overall, subjects exhibit a responsibility premium of 29%–48% when paying to repair personally caused harm. These results have important implications for the limits of individual responsibility in consumer markets, the demand for voluntary carbon offsets, and the efficiency of CSR activities.

1 Introduction

Many individuals and firms voluntarily incur private costs in order to mitigate negative externalities. Consumers regularly pay a premium to minimize their footprint: they pay a premium to offset their delivery emissions (Rodemeier, 2026) or for products that carry fair-trade labels (Hainmueller et al., 2015). When made aware of their impact on the environment, they substitute to a less carbon-intensive diet (Schulze-Tilling, 2025) or increase their donation to a charity fighting global warming (Semken, 2026). Likewise, and driven by stakeholder demand, firms go beyond their legal obligations and sacrifice short-term profits to avoid generating negative externalities (Bénabou and Tirole, 2010). For instance, Patagonia commits itself to produce their clothes with 100% organic cotton. Various outcome-based accounts (Becker, 1974; Fehr

[†]Department of Economics and Business, Universitat Pompeu Fabra (e-mail: horia.guias@upf.edu).
I am deeply thankful to Larbi Alaoui and Chiara Aina for their valuable comments and support. This paper also benefitted from discussions with Jose Apesteguia, Gaël Le Mens, Rosemarie Nagel, and Marta Serra-Garcia. I am also very thankful for the feedback received from audiences at FUR 2026, ESA 2026, NCBEE 2026, as well as the the 7th QMUL Economics and Finance Workshop for PhD & Post-doctoral Students. All remaining errors are my own.

and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002) and models of image concerns (Andreoni, 1990; Benabou and Tirole, 2006) have been proposed to account for these motivations to behave in a socially responsible manner –- being willing to sacrifice own payoff to increase welfare of others. What has received relatively little attention in economics, however, is how perceived responsibility for negative externalities drives behaviour.

The distinction between responsibility-based and outcome-based preferences is subtle – and matters for behaviour. To see it, consider the following example related to the growing market for voluntary carbon offsets: many individuals are willing to pay a price to offset their own carbon emissions. But they may not want to offset the same carbon emissions at the same price if they are caused by someone else. What varies in the two settings is the decision maker's perception of having caused harm to the passive other. Because anticipated outcomes are identical, outcome-based preferences cannot account for this behavioural difference. Instead, a responsibility-driven account explains why individuals want to offset their own but not others' emissions. This market is economically important: it has doubled its size since 2018 (Ecosystem Market Place, 2025) and is expected to reach a market volume of $50 billion by 2030 (Blaufelder et al., 2021).

In this paper, I study how individuals perceive their own causal responsibility for negative externalities. In particular, I examine how, shaped by previously caused harm, perceived responsibility drives behaviour as opposed to anticipated consequences. I develop a theoretical model in which, beyond other-regarding preferences based on final outcomes, a decision maker cares about maintaining a moral self-image that is activated by a backward-looking perception of responsibility: unrepaired personally caused harm adversely affects self-image while unrepaired other-caused harm does not. I show that the desire to restore a moral self-image drives socially responsible behaviour and may even override efficiency or equity considerations. However, responsibility limits the moral domain to personally caused harm. I provide a sharp revealed preference test that distinguishes responsibility-based preferences from outcome-based preferences. Based on the theoretical framework, I design an online experiment and structurally estimate preferences. I find considerable heterogeneity in the influence of perceived responsibility on behaviour: subjects exhibit a responsibility premium of 29%-48% when paying to repair personally caused harm.

To theoretically introduce the concept of backward-looking responsibility and understand its interaction with self-image concerns in the context of negative externalities, I study a decision maker in a stylized non-strategic setting where her actions generate a negative externality to a passive other, and she observes exogenous background harm. Crucially, the passive other takes no actions. To isolate responsibility-based preferences, I abstract away from strategic interactions by assuming that background harm is of unknown origin to the decision-maker. While she knows she did not cause this background harm, she may hold any other belief regarding its origin. She then makes two decisions in this framework. In the Externality-Creation Stage, she faces a binary choice whether to forego a personal benefit and personally contribute to the negative externality. In the following Reparation Stage, equipped with a linear reparation technology, she can repair the harm caused by herself and the background harm by paying a constant unit reparation cost that is given. In the Reparation Stage, I assume for tractability that the decision maker has linear other-regarding preferences and cares about maintaining a moral self-image. However, this self-image interacts with a backward-looking responsibility function that tracks any unrepaired harm that has been personally caused in the past. Because she perceives herself as responsible for the personally caused harm, leaving it unrepaired adversely affects her moral self-image. Conversely, she does not perceive herself as responsible for

the background harm, and thus leaving this type of harm unrepaired does not affect her moral self-image.

The predicted reparation choice depends on the monetary unit reparation cost. More specifically, the parameter space is partitioned into three regions. If the unit reparation cost is sufficiently low, the decision maker repairs total harm. If the unit reparation cost lies in an intermediate responsibility-based region, she repairs exclusively personally caused harm. In the responsibility-based region, pure outcome-based preferences no longer predict an own-payoff sacrifice to increase the passive other's payoff. However, because unrepaired personally caused harm also adversely affects her moral self-image, the decision maker is willing to pay a responsibility premium to repair personally caused harm. But as she does not perceive herself as responsible for background harm, she leaves it entirely unrepaired. Finally, if the unit reparation cost is sufficiently high, she does not repair any harm.

Based on the predictions of the theoretical framework, I propose a sharp revealed-preference test to infer the responsibility-based preferences I introduce from choices. To this end, and conditional on the decision maker having caused harm ex ante, I derive comparative statics with respect to both the unit reparation costs and the causal history of the negative externality. I measure the causal history by the decision maker's relative share of the total externality caused. While variation in this share changes the history of how the negative externality has been generated, it does not change the final consequences in the Reparation Stage. Hence, any class of preferences whose domain is defined over final consequences predicts behavioural invariance. In contrast, whenever the motivation to restore one's moral self-image overrides outcome-based concerns, the decision maker responds to variation in causal history.

I extend the theoretical framework by letting the effect of perceived responsibility on moral self-image interact with background harm. On the one hand, perceived responsibility can be attenuated by background harm. If the decision maker observes the harm personally caused coexisting with a large amount of background harm, she may feel less responsible, and hence her moral self-image is less affected. Therefore, she is less willing to sacrifice payoff to repair personally caused harm compared to the benchmark model. On the other hand, while diffusion of responsibility has been studied in economics, what has received relatively little attention is the *amplification* of responsibility. This happens if the decision maker does not observe personally caused harm in isolation but assumes partial responsibility for the total harm, acknowledging that her harm contributes to a more severe negative externality. This perception increases the effect of perceived responsibility on moral self-image, and she is willing to sacrifice more to repair personally caused harm compared to the benchmark model.

Based on the theoretical framework, I design an online experiment whose objective is to quantify the heterogeneity in the effect of perceived responsibility on preferences. Subjects make a series of decisions that have consequences for themselves and for a charity acting in the role of the passive other. In the Externality-Creation Stage, subjects choose whether to earn a personal benefit or not. If they earn money, a donation that funds the distribution of items by a charity is withdrawn. Consequently, these charity items are no longer distributed. In the Reparation Stage, subjects can spend their earnings to restore personally caused lost items, background lost items, or both. The experiment features two key orthogonal variations in the Reparation Stage. First, conditional on having generated harm in the first stage, I hold the total number of lost charity items constant and exogenously vary the causal share of harm caused by the decision maker. This variation isolates causal history from final consequences. Second, I exogenously vary the unit price of restoration to elicit subjects' willingness to pay to repair personally caused versus background harm. To

structurally estimate subject's preferences at the individual and aggregate level, I impose a linear utility function which allows for different weights on unrestored personally caused item loss and unrestored already lost items.

I find considerable heterogeneity in the effect of perceived responsibility on behaviour. Some subjects behave in a way that is consistent with outcome-based preferences. They sacrifice a fraction of their earnings to restore lost items regardless of how the loss has been generated. However, most subjects are driven by a perception of responsibility for the item loss they have caused in the past. Also, I find that around 40% of subjects do not want to cause harm even though it is less costly to restore personally caused lost items than avoiding item loss ex-ante. For the remaining 60% of subjects, I can exogenously vary the causal history of item loss, and I structurally estimate their preferences. At the individual level, I find that around 70% of subjects place a higher weight on unrepaired personally caused item loss, thus exhibiting a higher willingness to pay to restore personally caused lost items. The median subject is willing to sacrifice 34 cents to restore €1 of already lost charity items. Moreover, she is willing to pay a 38% responsibility premium to restore personally caused lost charity items. Finally, her perception of responsibility amplifies, thus suggesting that she may perceive herself as partially responsible for total item loss. Her responsibility amplifies locally at a rate of 4% per €1 of already lost items. However, there is heterogeneity in amplification: for 30% of subjects, perceived responsibility attenuates in the presence of already lost items.

At the aggregate level, I estimate finite mixture models to provide a parsimonious characterization of heterogeneity. I find that the influence of perceived responsibility on behaviour is pervasive across all identified preference classes in the representative agent model and a two-type model. All preference classes exhibit a statistically significant responsibility premium ranging between 29% and 48% when paying to restore already lost items. Perceived responsibility either diffuses or is not affected by already lost items; thus, the amplification effect exists for a minority but is rather weak.

The primary contribution of this paper is to theoretically formalize and empirically validate a mechanism through which backward-looking perception of responsibility drives restorative behaviour. From a theoretical framework, I derive a sharp revealed-preference test that distinguishes responsibility-driven self-image concerns from standard consequentialist and outcome-based preferences. The revealed-preference test informs the design of a laboratory experiment that allows me to cleanly isolate backward-looking perception of responsibility, empirically test the model's predictions and quantify its impact on economic decision-making. The mechanism can be extended to study the limitations of individual responsibility in consumer markets or to shed light on corporate social responsibility (CSR) practices of firms that are primarily driven by responsibility-based or outcome-based concerns.

As a second contribution, I provide granular behavioural evidence for perceived own responsibility for negative externalities that is possibly non-consequentialist. I design a laboratory experiment that isolates a backward-looking perception of responsibility. Then, I show that many subjects perceive themselves as responsible for the harm they have caused in the past. If the price to restore lost items is moderately high, they do not want to restore all items but those whose loss they have personally caused. Conversely, other subjects have a forward-looking perception of responsibility and regard the Reparation Stage as an opportunity to undo a negative externality, regardless of who has caused the loss of charity items.

The causal history of a negative externality matters for behaviour, in particular if individuals anticipate the

mechanism of how backward-looking responsibility affects their self-image.[1] Understanding this mechanism is crucial for market settings: in consumer markets, it explains why consumers pay a premium for green or plastic-free goods to neutralize their own footprint yet refuse to pay to offset carbon emissions or remove plastic waste generated by other actors. Because backward-looking responsibility restricts a consumer's moral domain strictly to personally caused harm, market activity leaves unaddressed a negative externality nobody feels responsible for. On the firms side, this mechanism sheds light on CSR practices, explaining why firm owners may want to go beyond their legal obligations and allocate resources toward remediating personally caused harm rather than maximizing global social impact elsewhere (Bénabou and Tirole, 2010). In labour markets, consider industries that generate negative externalities, such as the weapon or tobacco industry. Moreover, suppose that workers are replaced when rejecting employment. Workers with responsibility-based preferences care about the externality they personally generate and avoid employment in these industries, leading the least concerned workers to sort in. Despite causing personal harm, workers who have outcome-based preferences sort into these jobs, if they have the opportunity to substantially reduce the externality on the job.[2] Recent experimental labour markets provide evidence for sorting of the most *immoral* workers into jobs that create negative externalities (Schneider et al., 2024). In sum, understanding *how* agents internalize negative externalities when making decisions is crucial to understand market equilibria.

The activation mechanism proposed in my theoretical model can be applied to study demand for voluntary carbon offsets. In recent years, demand for voluntary carbon offsets has increased (Rodemeier, 2026) and is regarded as an efficient tool to abate carbon emissions (Calel et al., 2025). I provide an additional account for preferences to purchase voluntary carbon offsets. To what extent do investors pay for offsetting carbon emissions they have caused in the past? To what extent do they pay for maximizing social welfare when mitigating a negative externality? The revealed-preference test can be used to discriminate between these motivations and can be extended to more complex externality-generation mechanisms and uncertainty about individual impact.[3]

In sum, whether perceived responsibility is based on anticipated outcomes or personally caused harm has important consequences for individual behaviour and ultimately for market and policy outcomes. Following the literature review, the paper proceeds as follows: Section II presents the theoretical framework. Section III discusses the experimental design and Section IV discusses the econometric strategy for the structural estimation. Section V presents the results, and Section VI concludes.

Related Literature: The closest related paper is Jakob et al. (2017) who provides behavioural evidence for a preference to clean up personally caused harm that leads to inefficient outcomes. Subjects are paired and paid in function of the pair's performance in an effort task. In a between-subjects-design, they vary whether a team member or an unrelated third party created harm beforehand - a large number of chickpeas landing on the floor as they do not hit a bin - and both team members receive a payment for the effort tasks only if one team member cleans up the mess. They find that the fraction of subjects who prefer to clean up the mess

[1] The idea of responsibility activating a moral self-image is studied in other fields. Moral Philosophy identifies causal responsibility as an important factor for blameworthiness (Smith, 1759; Cicero, 1913; Halpern, 2016; Talbert, 2019). Social psychologists study how responsibility activates personal norms (Schwartz, 1977; De Groot and Steg, 2009) and moral psychologists provide suggestive evidence that it matters for behaviour in the trolley problem (Greene et al., 2009).

[2] For example, they could roll out responsible marketing campaigns such as avoid advertising cigarettes in the vicinity of schools.

[3] Rodemeier (2026) conducts a field experiment with a German online supermarket in which online shoppers only receive information about the average carbon emissions that correspond to one delivery. He shows that, conditional on purchasing, their marginal willingness to pay for marginal impact beyond the average carbon footprint associated to one delivery is 0.

by themselves is 30% higher if they have created the mess beforehand, even though delegation to the other team member is the efficient decision. I build on the idea in Jakob et al. (2017), by theoretically modelling such a responsibility-driven moral self-image.[4] Moreover, I design a novel experiment in which I isolate the causal history of a negative externality from final consequences in a reparation decision with which I can measure at the individual level the relative strength of outcome-based and responsibility-based preferences.

Gneezy et al. (2014) provides behavioural evidence for a conscience accounting hypothesis: following a moral transgression, subjects are more likely to donate a small amount to a charity than in the absence of a moral transgression. As an explanation, they propose a dynamic evolution of guilt paired with time-inconsistent preferences to make a donation. The ex-ante self prefers not to donate but, driven by guilt resulting from the moral transgression, the ex-post self wants to donate. I propose a mechanism in which individuals want to *exactly* repair a quantifiable harm they have previously caused and I do not assume time inconsistency in preferences.

A large literature shows how contextual features affect pro-social behaviour. Dana et al. (2007) were one of the first to show experimentally that pro-social motivations of individuals are not stable but context driven, especially when the relationship between a subject and final outcomes is not transparent. An economically relevant setting is the diffusion of responsibility in trading and consumer markets. A large and relatively young experimental literature studies how a market setting affects moral values relative to a setting of single decision-making (Falk and Szech, 2013; Bartling et al., 2015; Kirchler et al., 2016; Pigors and Rockenbach, 2016; Irlenbusch and Saxler, 2019; Bartling et al., 2019; Sutter et al., 2020; Falk et al., 2020; Bartling and Özdemir, 2023; Ziegler et al., 2024). Moreover, experimental evidence shows that perception of own responsibility is malleable and can be delegated on others (Hamman et al., 2010; Bartling and Fischbacher, 2012).

However, what has been commonly assumed in this literature is that perception of responsibility is purely forward-looking. I show that it can be backward-looking, and thus non-consequentialist. Moreover, this paper is the first in providing evidence that for some subjects their perception of responsibility amplifies, if they regard the personally caused lost items in a broader context among the already lost items, thus at least partially attributing responsibility for the total item loss to themselves.

A backward-looking perception of responsibility has been shown to influence redistribution by shaping which factors individuals are held responsible for (Cappelen et al., 2007, 2010, 2013a,b; Almas et al., 2020; Andre, 2025), personal production choices (choice egalitarianism), choices and skills (meritocracy), or all outcomes, including factors beyond individual control (strict libertarianism). Redistribution then occurs in proportion to factors others are held responsible for. Instead of focusing on the source of inequality I focus on the source of negative externalities. Moreover, I focus on self-attribution of causal responsibility for negative externalities.

Existing literature conceptualizes moral self-image either as a belief over one's core moral values (Bénabou and Henkel, 2025) or as an internalized norm (Akerlof and Kranton, 2000; Alger and Weibull, 2013; Ellingsen

[4]Hart and Zingales (2017) study how socially responsible owners vote whether a company should be dirty or clean. A dirty company makes higher profits but generates a negative externality. They model responsibility-driven preferences for the damage by assuming an owner feels responsible if she directly votes for a company going dirty but not if she abstains from voting that it goes green. However, they do not provide behavioural predictions if an investor does not perceive herself as responsible. I build on this idea by isolating a responsibility-driven from outcome-based social preferences and validating this preference in the lab.

and Mohlin, 2026). Brekke et al. (2003) study how the desire to maintain a moral self-image affects public good giving. While consumers in their framework want to view themselves as socially responsible, they propose a different self-image mechanism. They model consumers with Kantian preferences who evaluate their self-image by comparing the taken action against a moral ideal derived from maximising social welfare in the hypothetical case all consumers would adopt that action. Building on the norm-based approach, I model moral self-image as a direct psychological cost incurred from personally causing harm. However, not addressing harm caused by others does not adversely affect moral self-image. I show theoretically and experimentally how perceived responsibility affects this moral cost and how anticipating this mechanism shapes decision-making.

By structurally estimating preferences at both the individual level and the aggregate level using finite mixture models, this paper contributes to the literature that structurally identifies social preferences (Charness and Rabin, 2002; Andreoni and Miller, 2002; Fisman et al., 2007; Iriberri and Rey-Biel, 2011, 2013; Van Leeuwen and Alger, 2024; Alger and Rivero-Wildemauwe, 2024). While existing structural estimations focus primarily on outcome-based preferences, my framework expands this domain by including and structurally estimating non-consequentialist responsibility-based preferences.

2 Theoretical Framework

I use a theoretical framework to formalize a responsibility-based preference and to understand its interaction with outcome-based preferences. I do not specify the exact nature of the latter, but examples include social welfare preferences (Charness and Rabin, 2002) and inequity aversion (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). In the second part of this section, I use the results derived from the theoretical framework to propose a sharp revealed preference test to disentangle the formalized responsibility-based preference from any type of outcome-based preferences. In this setting, I define a decision maker's responsibility for a negative externality as her contribution to the negative externality that a passive other suffers.

2.1 Formalizing Responsibility-Based Preferences

Responsibility can affect decision making indirectly through outcome-based preferences or directly through responsibility-based preferences. To understand how own perceived responsibility is shaped by past actions versus anticipated outcomes, I study a decision maker (DM) in a stylized two-stage non-strategic setting. In the first stage, earning money generates a negative externality to a passive other. In the second stage, the DM can sacrifice her own payoff to repair for the harm caused to the passive other. Crucially, the passive other takes no action in this framework. Reparation means that the DM has access to a technology through which she can *undo* the externality previously caused.[5] The DM is fully informed about the magnitude of the negative externality generated by her actions. Moreover, actions map deterministically into outcomes, thus the negative externality does not arise as an exogenous shock.[6]

[5]Real-world examples include a consumer's decision to privately offset carbon emissions or a financial contribution to help removing industrial plastics from the ocean.

[6]This assumption rules out settings in which the DM causes a negative externality by accident such as dropping a piece of plastic to the ground. Under such a context, the DM's responsibility-based preferences may account for an ex-ante responsibility

To disentangle responsibility-based from outcome-based preferences, I impose the following assumptions. To rule out strategic concerns, the DM observes background harm $h_{\text{back}} \geq 0$ which she knows she did not cause, but other than that she is free to hold any belief regarding its source.[7] Background harm is an exogenous shortfall in the passive other's initial endowment ω_r. This assumption fulfils two more purposes: first, the responsibility function I formalize discriminates between the harm caused by the DM and the harm not caused by the DM, regardless of the latter's origin. Second, it allows me to disentangle a responsibility-based from an outcome-based preference by varying how a given negative externality has been generated. Finally, to rule out forward-looking stage-1 plans to repair harm, I assume the Reparation Stage arrives to the DM as an exogenous shock after she has made a choice in the Externality-Creation Stage.[8]

Formally, let ω_r be the passive other's finite initial endowment. The DM faces a binary choice to earn a monetary benefit $y \geq 0$ and generate additional harm $h_{\text{own}} \geq 0$. Let $b \in \{0, 1\}$ be the DM's choice in the Externality-Creation stage and $b = 1$ denote the decision to earn the monetary benefit. The DM's choice is mapped into final payoffs for her and the passive other in the following way: $x^{\text{EC}}_{\text{DM}}(b) = by$ and $x^{\text{EC}}_{r}(b) = \omega_r - h_{\text{back}} - bh_{\text{own}}$ define the DM's and passive other's final payoffs in the Externality-Creation stage, respectively. In the Reparation Stage, and equipped with a reparation technology with a constant monetary unit reparation $\kappa \geq 0$, she makes a one-shot decision to repair harm. She can choose any amount $c \in [0, H]$ that should be repaired, where $H = bh_{\text{own}} + h_{\text{back}}$ is the total harm caused to the passive other.[9] Finally, let $x^{\text{RS}}_{\text{DM}}(b, c) = x^{\text{EC}}_{\text{DM}}(b) - \kappa c$ and $x^{\text{RS}}_{r}(b, c) = x^{\text{EC}}_{r}(b) + c$ be the final payoffs of the DM and passive other in the Reparation Stage, respectively. I assume that $\omega_{\text{DM}} > \kappa H$ ensuring the DM always maintains a payoff advantage over the passive other.

The assumptions leave preferences over anticipated outcomes or over the harm caused in the past as the unique drivers of reparation of negative externalities. If own responsibility is entirely shaped by anticipated outcomes, then there is no need to formally model it. However, it might also be backward-looking, and thus not outcome-based. To this end, I formally introduce a backward-looking responsibility function. It examines the total negative externality caused to the passive other and tracks the unrepaired externality caused by the DM. Formally, I define the responsibility function as follows:

$$\rho(b, c; h_{\text{own}}) = \max\{bh_{\text{own}} - c, 0\} \tag{1}$$

The responsibility function has three key properties: first, it offsets the harm caused by the DM in the first stage with the repaired amount. Second, it attributes responsibility to the DM only for the unrepaired

and for an ex-post responsibility. Ex-ante, the DM did not want to cause harm but ex post she did. Thus, perception of responsibility may be subjective. By keeping the framework deterministic the ex-ante and ex-post perception of responsibility does not differ.

[7] Battigalli and Dufwenberg (2007) model guilt in strategic interactions. It arises when a player internalizes that her action fails to achieve the payoff expected by her opponent, hence her second-order belief. Their model does not yield a sharp prediction that the DM is expected to repair personally caused harm. In my model, guilt arises from violating an internalized personal norm that is shaped by the harm caused in the past. See Dillenberger and Sadowski (2012) and Noor and Ren (2023) for axiomatic approaches on shame and guilt in the context of violating a personal norm.

[8] For example, this rules out time-inconsistent preferences. The first-stage self of the DM wants to earn money and not repair anything. However, she is naïve and does not anticipate that the second-stage self suffers guilt and strictly prefers to repair harm. See Gneezy et al. (2014) for a similar form of time-inconsistent preferences.

[9] I do not model explicitly the DM's initial endowment. However, I assume that the DM is not budget constrained when repairing harm in the second stage. This assumption could be obtained by modelling an initial endowment ω_{DM} and assuming it is sufficiently high.

harm she has caused, and not for every change in the passive other's payoff caused by the DM's actions. Finally, it is backward-looking and discriminates between unrepaired harm caused by the DM and unrepaired background harm. Thus, the DM is responsible for unrepaired personally caused harm but not for unrepaired background harm.[10]

I assume that own responsibility for negative externalities adversely affects the DM's moral self-image. Let $\eta_{\text{resp}} \cdot \rho(H-c; bh_{\text{own}})$ denote her moral self-image. The higher the harm she has previously caused the worse the effect of her decision on her moral self-image.[11] When evaluating alternatives, the DM anticipates the harm she has caused, and thus the effect of a backward-looking responsibility on moral self-image. Therefore, I assume that responsibility-based preferences enter the DM's utility function as a cost. In addition to responsibility-based preferences, I allow the DM to have outcome-based concerns that enter as a benefit into her utility function.[12] In addition, the DM cares about her own payoff. Finally, for tractability, I assume that the outcome-based preferences are linearly separable from responsibility-based preferences. I define the DM's preferences as follows:

$$u(c; b, h_{\text{own}}, h_{\text{back}}) = x^{RS}_{\text{DM}}(b,c) + \eta_{\text{outcome}} \cdot x^{\text{RS}}_{r}(b,c) - \eta_{\text{resp}} \cdot \rho(b,c; h_{\text{own}}), \tag{2}$$

where the last term captures the moral cost proportional to the backward-looking perception of responsibility and the parameter $\eta_{\text{outcome}} \geq 0$ captures the strength of outcome-based concerns for the passive other.

The main result of this model is summarized in Proposition 1 and derives comparative statics results for the unit reparation cost κ.

Proposition 1. ***Comparative Statics in*** κ. *In any reparation situation the following is true:*

a) ***Purely Outcome-driven reparation:*** *if* $\kappa \leq \eta$, *then the DM repairs total harm* ($c^* = h_{own} + h_{back}$).

b) ***Purely Responsibility-Driven reparation:*** *if* $\eta < \kappa \leq \eta + \eta^{\text{i}}_{resp}$, *then the DM repairs exclusively personally caused harm* ($c^* = bh_{own}$).

c) ***No Reparation****: if* $\eta + \eta^{\text{i}}_{resp} < \kappa$, *then the DM does not repair any harm* ($c^* = 0$).

Figure 1 illustrates the optimal reparation choice c^* as a function of the unit reparation cost κ.

The unit reparation cost can be partitioned into three regions. If the unit reparation cost κ lies in the outcome-based region, the DM gains utility from repairing background harm on top of repairing personally

[10] Alternative mechanisms are imaginable: for instance, the DM may consider the generation of a negative externality a bad deed and mentally offsets this bad deed with a good deed. Thus, she exhibits a certain willingness to sacrifice to perform a good deed that, conditional on harm reparation, is insensitive to the amount of harm repaired.

[11] In terms of the framework developed by Ellingsen and Mohlin (2026), a backward-looking perception of responsibility activates a strict duty to repair personally caused harm. An alternative interpretation is to call $\rho(H-c; bh_{\text{own}})$ the DM's causal responsibility and $\eta^{\text{i}}_{\text{resp}}\, \rho(H-c; bh_{\text{own}})$ her moral responsibility.

[12] I do not make specific assumptions on the exact nature of outcome-based preferences. These concerns could reflect altruism (Becker, 1974; Charness and Rabin, 2002), or aversion to advantageous equity concerns (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). By assumption, I rule out aversion to disadvantageous inequality.

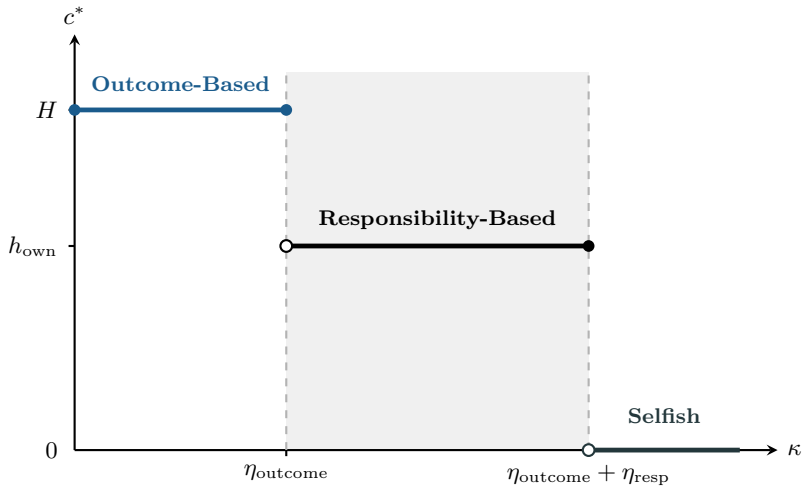


Figure 1: Optimal reparation c^* as a function of unit reparation cost κ (Proposition 1).

caused harm. Thus, optimal reparation c^* is located at the upper corner, and is primarily driven by outcome-based preferences given that responsibility-based preferences predict a zero willingness to sacrifice to repair background harm. If κ lies in the responsibility-based region, responsibility-based preferences *override* outcome-based preferences and the DM repairs personally caused harm h_{own}. Under pure outcome based preferences ($\eta_{\text{resp}} = 0$) the DM no longer gains utility from repairing harm. However, under responsibility-based preferences, she gains additional utility from restoring her moral self-image. Therefore, if $\eta_{\text{outcome}} < \kappa \leq \eta_{\text{outcome}} + \eta_{\text{resp}}$, the DM is willing to pay a responsibility premium to repair personally caused harm. Crucially, this responsibility premium is entirely driven by responsibility-based preferences. However, because responsibility-based preferences ignore background harm, the latter remains unrepaired.

Appendix A.1 shows the proof of Proposition 1.

2.2 Revealed Preference Test: Disentangling Outcome-Based and Responsibility-Based Preferences

Proposition 1 establishes that when the unit reparation cost falls in the responsibility-based region ($\eta < \kappa \leq \eta + \eta_{resp}$), responsibility-based preferences override outcome-based concerns. In this region, a purely outcome-driven agent would choose no reparation ($c^* = 0$), whereas an agent driven by backward-looking responsibility derives utility from restoring her moral self-image and exclusively repairs personally caused harm ($c^* = h_{own}$). However, to disentangle responsibility-based preferences as defined in equation (2) from any outcome-based preferences, there must exist choices that can be rationalized by a utility function as given in equation (2) but cannot be rationalized by *any* outcome-based preference.

To derive a sharp revealed preference test that disentangles responsibility-based preferences from outcome-based preferences consider the two-stage framework as defined in the previous subsection and fix a total harm that has been caused $\bar{H} = h_{\text{own}} + h_{\text{back}}$. Moreover, assume the DM earns the monetary benefit in stage 1. I then define the DM's relative share on $\bar{H}$ by $\gamma \in [0, 1]$, thus the DM's absolute share on $\bar{H}$ is given by $\gamma\bar{H}$ and the absolute share of background harm by $(1-\gamma)\bar{H}$. I measure the causal history of $\bar{H}$ by the DM's

relative share γ. Notice that final stage-2 outcomes do not depend on γ and are given by $x_{\text{DM}}^{\text{RS}} = by - \kappa c$ and $x_r^{\text{RS}} = \omega_r - (\bar{H} - c)$ for the DM and passive other, respectively. Consequently, any purely outcome-based utility function $U(x_{DM}^{\text{RS}}, x_r^{\text{RS}})$ is invariant to changes in γ. The reason is that variation in the causal history of $\bar{H}$ is decoupled from final stage-2 outcomes and only the composition of $\bar{H}$ varies. But outcome-based preferences do not care about how $\bar{H}$ has been generated. Therefore, they cannot predict a response to variation in γ. Instead, if $\kappa \in (\eta_{\text{outcome}}, \eta_{\text{outcome}} + \eta_{\text{resp}}]$, then reparation is driven by responsibility-based preferences. Because reparation tracks personally caused harm unit by unit, it will increase in response to an increase in the DM's relative share γ. Thus, the variation in causal history of $\bar{H}$ produces a choice that can be rationalized by a responsibility-based preferences but by *no* outcome-based preferences. Proposition 2 formalizes this result.

Proposition 2. ***Sharp Revealed Preference Test: Disentangling Responsibility-Based from Outcome-Based Preferences*** *Let $x_{DM} = x_{DM}^{RS}(b, c)$ and $x_r = x_r^{RS}(b, c)$. Suppose the DM is unobserved by others, observes exogenous background harm of unknown origin (h_{back}) and cannot plan reparation choices in the first stage. Moreover, fix a total negative externality $\bar{H} = h_{own} + h_{back}$ and suppose the DM takes the monetary benefit in stage 1. Then the following is true in the second stage:*

1) *No strictly consequence-based utility function, $U(x_{DM}, x_r)$, can rationalize $\frac{\partial c^*}{\partial \gamma} > 0$.*

2) *Suppose the DM has both outcome-based and responsibility-based preferences as defined in Equation (2). Then if $\eta_{outcome} < \kappa \leq \eta_{outcome} + \eta_{resp}$, her reparation choice responds to an in crease in the relative share γ, $\frac{\partial c^*}{\partial \gamma} = H > 0$. In particular, she will repair exclusively for personally caused harm. Otherwise, $\frac{\partial c^*}{\partial \gamma} = 0$.*

This sharp revealed preference test for responsibility-based concerns first underlines that *no* outcome-based preference can rationalize tracking and reparation of personally caused harm because it directly produces a preference reversal between the two chosen alternatives by the DM. The second part relies on the results from Proposition 1, showing that by allowing for responsibility-based preferences as defined in (2) the DM responds to an increase in her relative share on $\bar{H}$.

Thus, under the identifying assumptions the sharp revealed preference test can be applied in the following way: fix a total harm that has been caused to a passive other $\bar{H}$. Then vary the unit reparation cost κ to trace out the responsibility-based region. Finally, vary the DM's relative share on $\bar{H}$ in order to produce choices that can be rationalized by responsibility-based preferences but not by outcome-based preferences. The experimental design will rely on this sharp revealed preference test. Appendix A.2 contains the proof of Proposition 2.

2.3 Attenuation and Amplification of Responsibility

The DM's preferences can be extended by allowing backward-looking responsibility to interact with the environment. In particular, I assume that it can attenuate or amplify in the presence of background harm. The DM's perceived responsibility could attenuate if she observes personally caused harm h_{own} among background harm h_{back}, especially if the latter is much higher. In other words, her responsibility diffuses, and hence the effect on moral self-image decreases with background harm. However, and what has received

relatively little attention in economics, is that responsibility may amplify with background harm. This occurs when the DM acknowledges that personally caused harm contributes to a much larger negative externality, thus partially assuming responsibility for total harm $H = h_{\text{own}} + h_{\text{back}}$ caused. Hence, the effect of responsibility on moral self-image increases in background harm. To parsimoniously capture both types of effects, I introduce an attenuation function $\delta(h_{\text{back}}; \alpha) = \exp(\alpha h_{\text{back}})$ where $\alpha \in \mathbb{R}$ is a diffusion parameter. $\alpha < 0$ captures responsibility attenuation while $\alpha > 0$ captures responsibility amplification as h_{back} increases. The specification of the utility function including the diffusion function is then given as follows:

$$u(c; bh_{\text{own}}, h_{\text{back}}) = x^{\text{RS}}_{\text{DM}}(b, c) + \eta_{\text{outcome}} x^{\text{RS}}_{r}(b, c) - \eta_{\text{resp}} \delta(h_{\text{back}}; \alpha) \rho(H - c; bh_{\text{own}}) \tag{3}$$

The utility function remains quasiconcave for all values of α. The main insights from Proposition 1 carry over as is shown in Proposition 3:

Proposition 3. *In any reparation situation where preferences are given by (3). The following holds:*

a) ***Comparative Statics in*** κ*.*

a1) If $\kappa \leq \eta$*, the DM repairs total historical harm* ($c^* = H$)*.*

a2) If $\eta_{outcome} < \kappa \leq \eta_{outcome} + \delta(h_{back}; \alpha)\eta_{resp}$*, the DM repairs exclusively own past harm* ($c^* = bh_{own}$)*.*

a3) If $\kappa > \eta_{outcome} + \delta(h_{back}; \alpha)\eta_{resp}$*, the DM does not compensate* ($c^* = 0$)*.*

The proof of proposition 3 is shown in Appendix A.3. In comparison to the benchmark model of outcome-based and responsibility-based preferences, under diffusion and amplification of moral self-image the responsibility region of κ in which perceived responsibility is the unique driver of reparation choices, shrinks and increases, respectively. This is illustrated in Figure 2 :

The main insights from the sharp revealed preference test carry over to the extended model. The only difference is the need to account for the interaction effect when varying γ. Under amplification of responsibility, the size of the responsibility-based region is higher compared to the benchmark model where $\alpha = 0$. However, an increase in γ causes background harm to decrease, and hence the amplification effect becomes smaller. Thus, an increase in γ causes the responsibility-based region to shrink. If responsibility attenuates, then the size of responsibility-based region is smaller relative to the benchmark model. However, increasing γ leads to a decrease in background harm and thus to a lower attenuation of responsibility. Thus, an increase in γ causes the size of the responsibility-based region to increase. In Appendix A.3 the proposition and the proof of the sharp revealed preference test are fully formulated.

In sum, the theoretical framework shows that accounting for the causal history of a negative externality is important to explain behaviour as it shapes the perceived responsibility of a DM. If responsibility drives behaviour, then the moral domain is limited to personally caused harm and she ignores background harm.

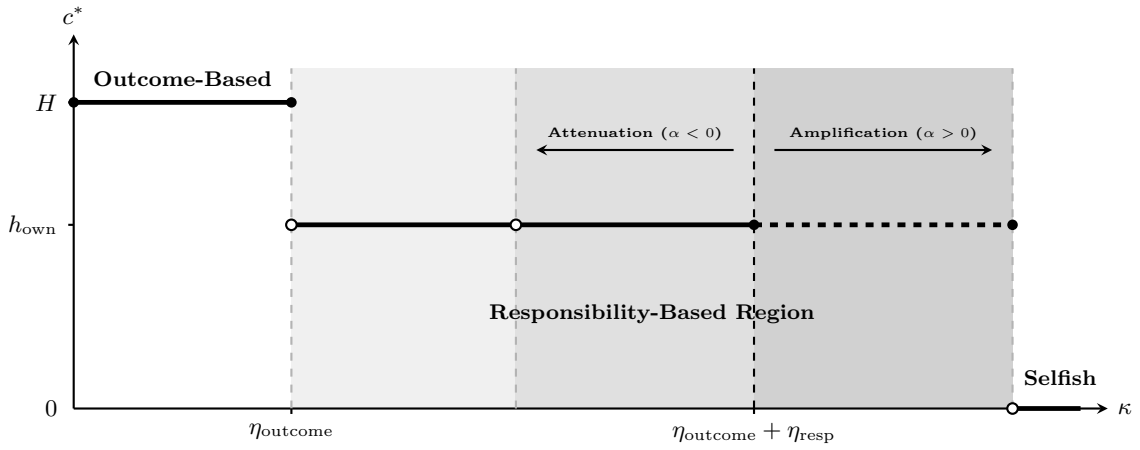


Figure 2: Optimal reparation c^* under responsibility attenuation ($\alpha < 0$) and amplification ($\alpha > 0$) as a function of unit reparation cost κ (Proposition 3).

The sharp revealed preference test derived from the theoretical framework informs the design of the lab experiment. The next section presents the experimental design in more detail.

3 Experimental Design

The main objective of the experiment is to measure the intensity of outcome-based and responsibility-based preferences for the charity's well-being. By relying on the sharp revealed preference test that I propose in the theoretical framework, I can disentangle responsibility-based from outcome-based preferences. Moreover, based on Proposition 1, varying the unit reparation cost allows me to measure the intensity of linear outcome-based and responsibility-based preferences. This section discusses in detail first the design elements that map the comparative statics exercise from the sharp revealed preference test proposed in Proposition 2 into an experimental setting. After that, it discusses the design elements that are needed to measure the relative strength of linear responsibility-based and outcome-based preferences. Finally, it discusses the implementation of the experiment.

3.1 First Stage: Generation of Negative Externalities

Participants were informed that their decisions had consequences for themselves as well as for the donation made to another charity. In order to ensure that participants do not expect any form of punishment or rewards from the passive other, I opted for a charity in the role of the passive other. To maximise the probability that the participant cares about the donation made the charity, they were given a charity selection task before receiving any instructions about the experiment. In this charity selection task, participant could select the recipient of a donation among a list of three charities. Because participants were Spanish residents, I focused on charities that are operating in Spain, and hence the participants are likely to be familiar with them. The donation made to these charities finances the provision of certain charity items. Table 1 presents the list of charities, the number of charity items financed by the donation, and the country in which the charity is

operative.

Table 1: List of charitable organizations that could receive a donation of €7 in a charity selection task. The selected charity acted in the role of the passive other throughout the experiment. In each session, a subject was selected at random for whom the donation was carried out.

Charity	**Social Issue**	**Charity Items Financed by € 7 Donation**	**Geographical Area**
Plant-for-the-Planet	Global Warming	Planting one tree	Peruvian Amazonas
Fundación Educo	Child Poverty	Distribution of two meals	Spain
Doctors without Borders	Inadequate Medical Care Access	Distribution of one vaccination kit (=14 vaccines) against Meningitis	Central Africa

The selected charity remains in the role of the passive other in all decision situations a participant faces in this experiment. In the first stage of the experiment, a participant faces 7 decision situations. In each of the 7 decision situations, the selected charity starts with a donation it is assigned to receive. Eventually, the participant will observe that part of the donation has been reduced. While she knows the amount of the donation that has been cancelled, she does not know its origin except that it is unrelated to the decision she makes in the first stage of the experiment. The participant is offered a monetary benefit. In case she accepts it, as an additional consequence, the remainder of the donation is cancelled, and thus the selected charity receives no donation. If the participant rejects the monetary benefit, the part of the donation that is not already cancelled remains intact. The donation that remains intact is carried out by the experimenter if the selected decision situation is selected for payment.[13] In order to make the negative externality salient to participants, it is not framed as a cancellation of the donation but in terms of the corresponding loss in charity items.[14] Figure 3 shows a screenshot of a decision screen a participant observed in the first part of the experiment. In order to incentivize participants to earn the monetary benefit, they were informed that they could spend the monetary earnings to restore the total number of items that have been lost.[15]

In the theoretical framework, I measure the causal history of the negative externality by the decision maker's relative share on the exact same negative externality that has been caused in the first stage. To isolate the causal history of the negative externality from the final outcomes of reparation choices in the second stage, the participant's relative share on the cancelled donation needs to vary in the experiment. This requires the participant to accept the monetary benefit in the first-stage decision situation. Then, the exact same donation cancellation can be generated in different ways. I design three pairs of decision situations in the first stage in which I fix the monetary benefit as well as the total amount of the cancelled donation but vary the participant's relative share. In two pairs the donation cancellation has a single source. In decision situations

[13] See the logistics of the experiment for more details on the payment.

[14] Experimental designs represent a negative externality in different ways. For example, in the experimental design of Bartling et al. (2015) the negative externality corresponds the reduction in payoff of a passive participant. Some experimental designs represent the negative externality as a cancellation of a generic donation to a charity (e.g. Falk, Neuber and Szech (2020). I follow recent experimental designs in which the negative externality is framed to participants as a loss in provided charity items such as the provision of school lunches, vaccines against Measles or the medical treatment against Tuberculosis ((Irlenbusch and Saxler, 2019; Sutter et al., 2020; Bénabou et al., 2024)).

[15] Participants did not know how much they had to pay to restore the lost items, and thus could make any intertemporal decisions.

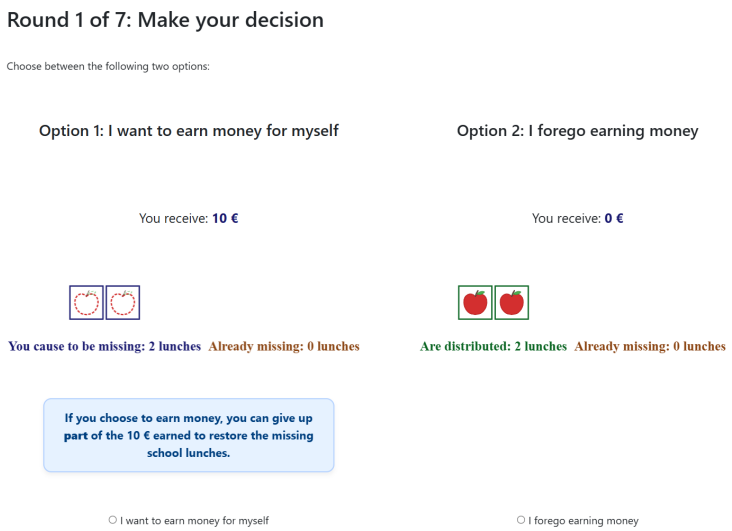


Figure 3: Decision Screen in the Externality Creation Stage

FullParticipant7 and **FullBackground7** the same cancelled donation of €7 is either fully participant-caused or fully exogenous. In the latter decision situation, earning a monetary benefit did not cause any additional cancellation in the donation. In **FullParticipant14** and **FullBackground** the cancelled donation of €14 is either fully participant-caused or exogenous. In one pair of decision situations, the total amount of the cancelled donation is €21 and of mixed source. In **ParticipantLow**, the share of the participant-caused donation cancellation is 1/3 and in **ParticipantHigh** it increases to 2/3. Finally, in decision situation **Symmetric** the share of the participant-caused donation cancellation is 1/2.[16] Figure 4 illustrates the main variation in the first stage of the experiment. While the monetary benefit is fixed within each pair decision situations in which the participant's relative share on the exact same total donation cancellation varies, I adjust the monetary benefit to the size of the total donation cancellation such that subjects are not constrained in the second stage when they choose to recover the cancelled donation.

3.2 Second Stage: Recovery of Cancelled Donation

For each first-stage decision situation in which the participant earned money for herself, she faced a set of reparation decisions in which she could sacrifice her own payoff to recover (parts of) the total donation that has been cancelled.[17] In the first-stage decision situations in which the participant earned money and the total donation cancellation is either fully participant-caused or exogenous, the participant faced a sequence of binary choices between recovering the total cancelled donation or not. In all other first-stage decision situations in which the participant earned money she could choose among four options: recover the total cancelled donation, recover the personally caused cancellation, recover the exogenous cancellation or

[16] To frame the negative externality in terms of the corresponding charity items that are lost, I use multiples of 7 in the variation of the total donation cancellation.

[17] If the participant did not earn money for herself, she did not have to repair for the exogenous harm she observed in order to minimize her time spent on the experiment.

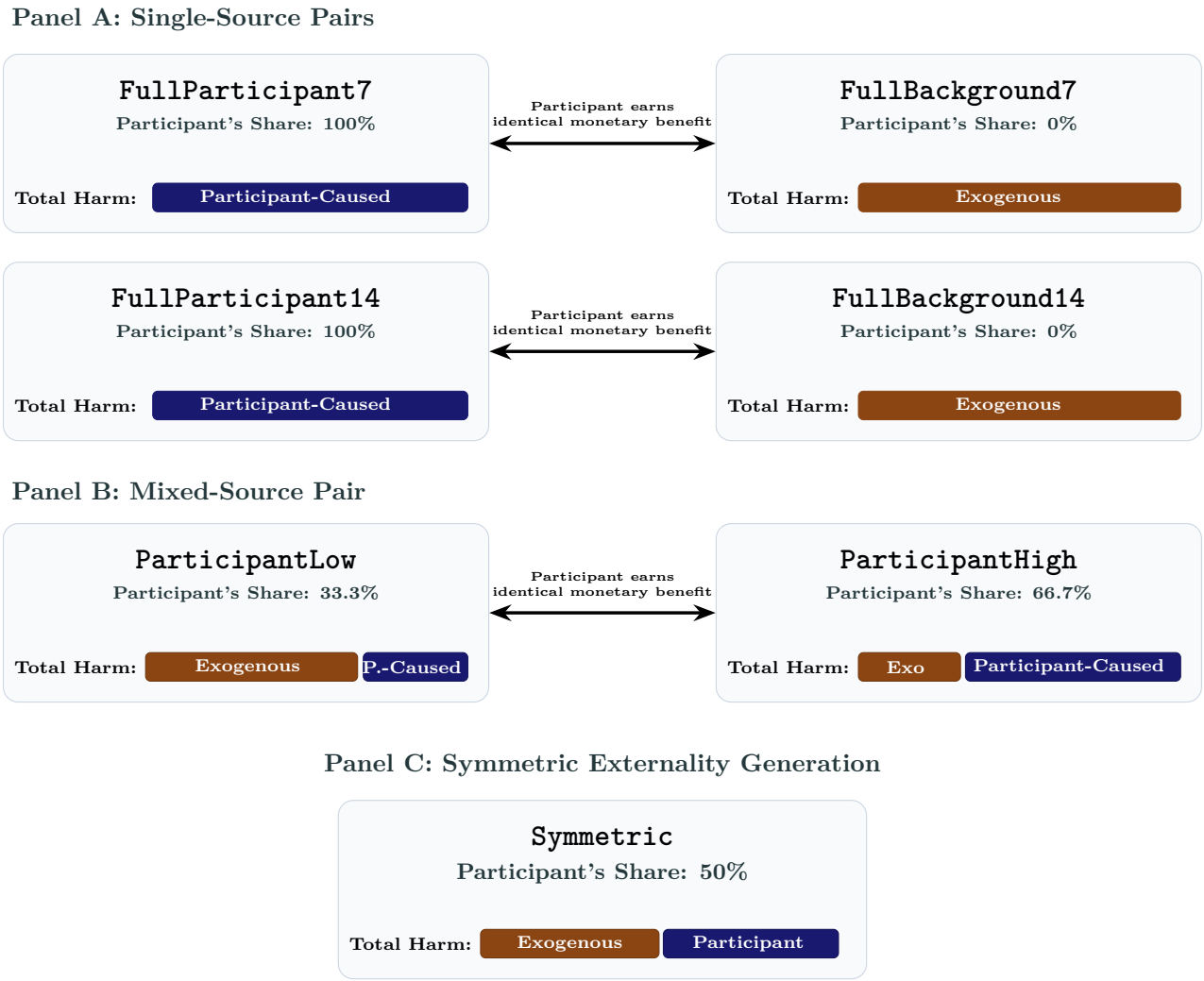


Figure 4: **First Stage Decision Situations**: This figure illustrates the main experimental variation in the first stage of the experiment across the 7 decision situations which is conditional on the participant earning the monetary benefit. In the pair of decision situations in Panel A, the same total donation cancellation is generated and it is either fully participant-caused or exogenous. In Panel B, the total donation cancellation consists of an exogenous cancellation as well as a participant-caused donation cancellation. The participant's relative share on the exact same cancelled donation is 1/3 and 2/3, respectively. In Panel C, the participant's relative share on the total donation cancellation is 1/2.

recover no cancelled donation.[18] The choice alternatives were framed in the corresponding number of lost charity items. Before making their reparation decision, participants were reminded about the total donation cancellation and its composition into a fraction caused by the participant and a fraction that is exogenous. Figure 5 illustrates a decision screen that was shown to participants.

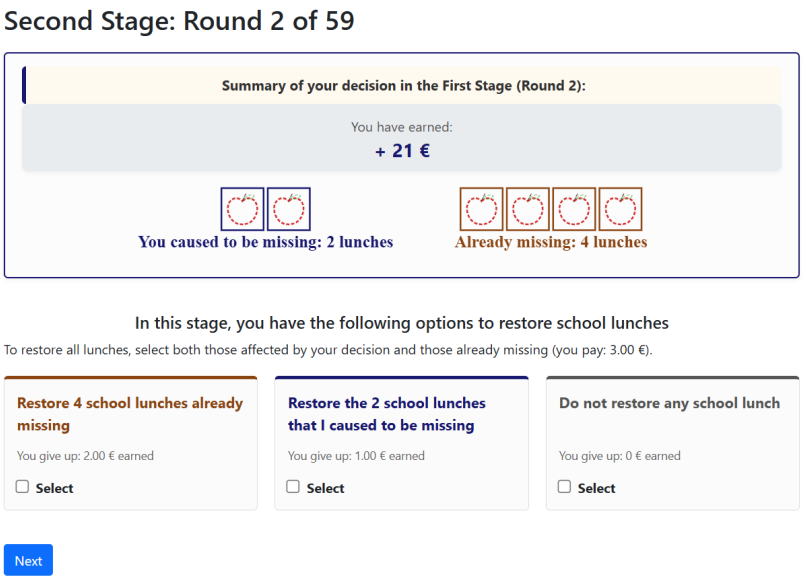


Figure 5: Decision Screen in the Reparation Stage

The main objective of the second stage is to measure the relative strength's of a participant's responsibility-based and outcome-based preferences for the donation made to the charity. I impose a linear structural utility function in which I allow a participant i to put different weights on the donation whose cancellation she has personally caused and on the exogenous donation. Moreover, I assume that the participant cares about her own monetary earnings. The linear structural utility function is defined as follows:

$$U_i(x_i, \tilde{h}^i_{\text{own}}, \tilde{h}^i_{\text{back}}) = x_i - \eta^i_{\text{own}} \cdot (1 + \alpha^i \tilde{h}_{\text{back}}) \cdot \tilde{h}_{\text{own}} - \eta^i_{\text{back}} \cdot \tilde{h}_{\text{back}} \tag{4}$$

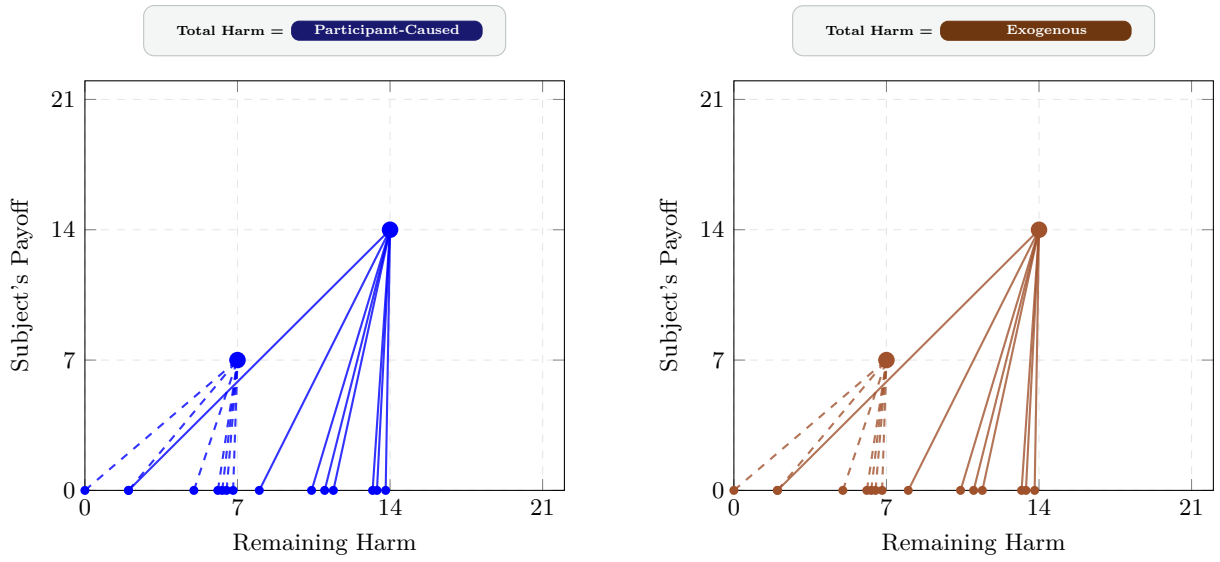
First, suppose $\alpha^i = 0$. Participant i is places weights η^i_{own} and η^i_{back} on the unrestored donation she personally cancelled and the unrestored exogenously cancelled donation, $\tilde{h}_{\text{own}}$ and $\tilde{h}_{\text{back}}$,respectively. She suffers a utility cost the higher the cancelled donation that is not restored. Under purely outcome-based preferences, she does not account for total donation cancellation was generated, and thus she treats the unrestored personally cancelled and unrestored exogenously cancelled donation as perfect substitutes ($\eta^i_{\text{own}} = \eta^i_{\text{back}}$). In contrast, under responsibility-based preferences, the marginal willingness to restore €1 of the personally cancelled donation receives a weight from outcome-based and responsibility-based preferences, and thus $\eta^i_{\text{own}} > \eta^i_{\text{back}}$.

α^i captures the percentage rate of change of the marginal willingness to restore €1 of the personally cancelled

[18] As I will structurally estimate linear preferences, I discretized the set of alternatives.

donation as background harm increases by €1. If $\alpha^i > 0$, then the effect of perceived responsibility on behaviour amplifies in the presence of the exogenously cancelled donation, and thus participant i's marginal willingness increases. In contrast, if $\alpha^i < 0$, then participant i's marginal willingness decreases in the presence of the exogenous cancellation, capturing the attenuation of perceived responsibility in the presence of the exogenously cancelled donation.

In total, there are three parameters that have to be estimated. Figure 6 represents the reparation choices for the donation cancellation that is either fully caused by the participant or is fully exogenous. On the y-axis I plot the participant's final payoff and on the x-axis the final donation that remains cancelled. One line segment represents a reparation decision and the two dots it connects represents the two final alternatives. Thus, the slope of this line segment can be interpreted as the unit reparation cost to recover €1 of the donation that has been cancelled. The parameter η^i_{own} is identified by the unit reparation cost κ that makes participant i indifferent between recovering the cancelled donation fully caused by herself and not recovering it. For all values of $\kappa < \eta^i_{\text{own}}$, participant i strictly prefers to recover the donation that has been fully cancelled by herself and for all values of $\kappa > \eta^i_{\text{own}}$ she strictly prefers not to recover it. The identical logic holds for the identification of η^i but for the reparation decisions in which the donation cancellation is fully exogenous. Moreover, as shown in Figure 6, the reparation decision in which the donation cancellation is fully participant-caused and fully exogenous differ only in the source of the donation cancellation and the final outcomes of the reparation decisions are identical.



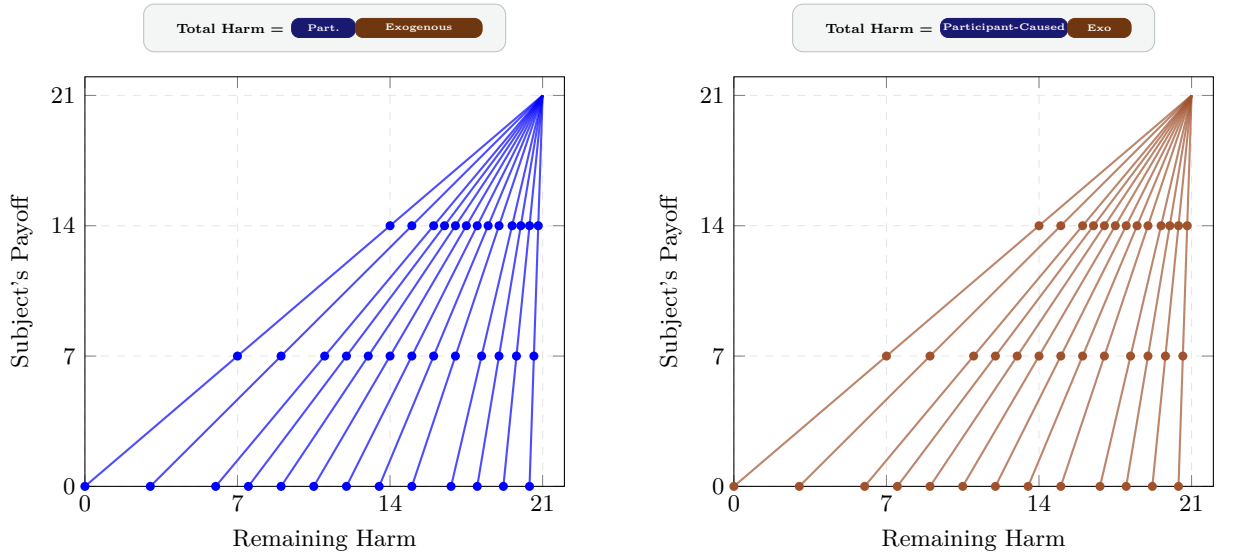
(a) Panel A: Cancelled Donation Fully Participant-Caused

(b) Panel B: Cancelled Donation Fully Exogenous

Figure 6: **Reparation Decisions in Stage 2:** Panel A shows reparation decisions for the cancelled donation that is fully participant-caused. Panel B shows the reparation decisions for the exact same cancelled donations that are fully exogenous. A line segment represents a reparation decision and a dot a final alternative. The slope of the line segment corresponds to the unit price to recover €1 of the cancelled donation.

Conditional on identification of η^i_{own} and η^i_{back}, the amplification parameter α^i is identified in the reparation decisions in which the donation cancellation has been generated in `ParticipantLow`. To see this, define

$\tilde{\eta}^i_{\text{own}} = \eta^i_{\text{own}} \cdot (1 + \alpha^i h_{\text{back}})$ and suppose $\tilde{\eta}^i_{\text{own}}$. Then, by relying on the logic of Proposition 1, it can be shown that if $\kappa = \tilde{\eta}^i_{\text{own}}$, the participant is indifferent between recovering the personally cancelled donation and not recovering any donation. In particular, if $\eta^i_{\text{back}} < \kappa < \tilde{\eta}^i_{\text{own}}$, participant i chooses to recover the personally cancelled donation and if $\kappa > \tilde{\eta}^i_{\text{own}}$, she chooses to not recover any cancelled donation.



(a) Panel A: Participant's Share 1/3 (`ParticipantLow`) (b) Panel B: Participant's Share 2/3 (`ParticipantHigh`)

Figure 7: **Reparation Decisions in Stage 2 (`ParticipantLow` vs. `ParticipantHigh`):** Panel A shows reparation decisions for decision situations with a lower share of participant-caused harm (`ParticipantLow`). Panel B shows reparation decisions for decision situations with a higher share of participant-caused harm (`ParticipantHigh`). A line segment represents a reparation decision and a dot a final choice alternative. The slope of the line segment corresponds to the unit price to recover €1 of the cancelled donation.

I create identical reparation choices for the donation cancellation generated in **`ParticipantLow`** and **`ParticipantHigh`** in the first stage. Thus, any purely outcome-based preference predicts behavioural invariance in the two sets of reparation decisions.

3.3 Baseline Logistics

The experiment has been conducted online on Prolific with the Spanish sample and the main sessions have been carried out in August 2026. In total, 168 subjects participated in the experiment. The median completion time was $\sim$ 37 minutes. The main analysis of the experiment is pre-registered and the experiment received Ethics approval from Universitat Pompeu Fabra.[19]

The experiment consists of 3 main parts. After receiving a set of detailed instructions explaining the role the charity and the decisions in the first stage, participants made 7 decisions. After that, they moved to Part II in which they received an additional set of instructions explaining them the reparation decisions. After completing Part II, subjects answered a brief questionnaire.

[19]The RCT ID is AEARCTR-0019297. Whenever there exists a part of the main analysis that has not been pre-registered I will point it out.

In order to incentivize choices one round from the first stage was randomly selected and two scenarios were possible. If the participant did not earn money in the randomly selected decision, then payment was based on the final outcomes after the first stage. Otherwise, the payment was based on one of the randomly selected reparation decisions in the second stage.

4 Econometric Strategy

I structurally estimate preferences using the stage-2 choices of subjects who accept the monetary benefit in the first stage, and thus for whom I can vary the causal history of a negative externality holding fixed outcomes in the Reparation Stage. This section presents the econometric strategy in detail.

4.1 Structural Utility and Random Utility Model

I assume subject i evaluates utility with error. Let $\epsilon_{i,\mathbf{x}_d}$ be the random component of choosing $\mathbf{x}_d \in X_d$, distributed i.i.d. Type-I Extreme Value across subjects and alternatives with scale parameter μ^i. I assume that a final alternative consists of a final payoff for participant x_i, an unrecovered participant-cancelled donation $\tilde{h}_{\text{own}}$ and an unrecovered exogenously cancelled donation $\tilde{h}_{\text{back}}$, i.e. $\mathbf{x}_d = (x_i, \tilde{h}_{\text{own}}, \tilde{h}_{\text{back}})$. Defining $\theta^i = (\eta^i_{\text{own}}, \eta^i_{\text{back}}, \alpha^i, \mu^i)$ as the vector of individual-specific parameters, the Random Utility Model (RUM) is:

$$\tilde{U}(\mathbf{x}_d; \theta^i) = U(\mathbf{x}_d; \eta^i_{\text{own}}, \eta^i_{\text{back}}, \alpha^i) + \epsilon_{i,\mathbf{x}_d} \tag{5}$$

Under these assumptions, the choice probability that subject i selects alternative $\mathbf{x}_d \in X_d$ in decision d is given by the multinomial logit formula:

$$\Pr(C_i(d) = \mathbf{x}_d; \theta^i, X_d) = \frac{\exp\{\mu^i U(\mathbf{x}_d; \eta^i_{\text{own}}, \eta^i_{\text{back}}, \alpha^i)\}}{\sum_{\mathbf{x}'_d \in X_d} \exp\{\mu^i U(\mathbf{x}'_d; \eta^i_{\text{own}}, \eta^i_{\text{back}}, \alpha^i)\}} \tag{6}$$

where $C_i(d)$ denotes the chosen alternative. The scale parameter μ^i captures choice sensitivity to utility differences: if $\mu^i = 0$, choices are completely random $(1/|X_d|)$, while as $\mu^i \to \infty$, choice probabilities converge deterministically to utility maximization.

Let $C_i = (C_i(1), \ldots, C_i(D))$ be the sequence of choices made by subject i across available choice sets $X_i = (X_1, \ldots, X_D)$. Subject i's individual likelihood contribution is:

$$f(\theta^i; C_i, X_i) = \prod_{d=1}^{D} \prod_{\mathbf{x}_d \in X_d} \Pr(C_i(d) = \mathbf{x}_d; \theta^i, X_d)^{I(C_i(d) = \mathbf{x}_d)} \tag{7}$$

where $I(\cdot)$ is an indicator function equal to 1 if subject i chose alternative $\mathbf{x}_d$ in decision d.

4.2 Aggregate Estimation using Finite Mixture Models

Estimating preferences at the individual level comes with drawbacks. First, using only 58 choices yields parameter estimates heavily driven by individual noise patterns (Jagelka, 2024). Second, estimating $168 \times 4 = 672$ parameters risks overfitting if the data is generated by a smaller set of preference types. Finally, characterizing population heterogeneity yields sharper policy and market insights.

I adopt Finite Mixture Models (FMM) to address these concerns. The population consists of K discrete preference types (θ_k, μ_k) with prior probabilities s_k where $\sum_{k=1}^{K} s_k = 1$. Let $\Phi = (\theta_1, \ldots, \theta_K, \mu_1, \ldots, \mu_K, s_1, \ldots, s_{K-1})$ denote the vector of parameters to be estimated. Subject i's likelihood contribution is:

$$m(\Phi; X_i, C_i) = \sum_{k=1}^{K} s_k f(\theta_k, \mu_k; X_i, C_i) \tag{8}$$

Using the estimated parameters $\hat{\Phi}$, Bayes' rule yields the ex-post probability τ_{ik} that subject i belongs to class k:

$$\tau_{ik} = \frac{\hat{s}_k f(\hat{\theta}_k, \hat{\mu}_k; X_i, C_i)}{\sum_{q=1}^{K} \hat{s}_q f(\hat{\theta}_q, \hat{\mu}_q; X_i, C_i)} \tag{9}$$

To select the optimal number of classes K, I follow standard literature (Bruhin et al., 2019) and minimize the Normalized Entropy Criterion (NEC):

$$\text{NEC}(K) = \frac{E(K)}{L(K) - L(1)} \tag{10}$$

where $E(K) = -\sum_{k=1}^{K} \sum_{i=1}^{N} \tau_{ik} \ln(\tau_{ik})$ measures classification entropy. The NEC balances improvements in model fit $(L(K) - L(1))$ against ambiguity in class assignment.

4.3 Individual-Level Estimations

Finally, I estimate individual-level parameters θ^i separately to report summary statistics on the empirical parameter distributions, without making inferences on individual statistical significance (Bruhin et al., 2019).

5 Results

The results analyse to what extent responsibility for negative externalities drives behaviour. I first discuss observed choice data before turning to the results of the structural estimation. Finally, I will compare the value added of allowing for responsibility-based preferences in the preference estimation. The analysis focuses on the selected sample that accepted the monetary benefit in the first part (113 subjects).

5.1 Choice Distributions: `FullOwn` vs. `FullOther`

Figure 8 illustrates the average reparation probabilities. The blue line represents the average reparation share in function of the unit reparation cost κ in `FullOwn` while the red graph plots the average reparation probability in function of κ in `FullBackground`. Each dot represents one reparation decision.

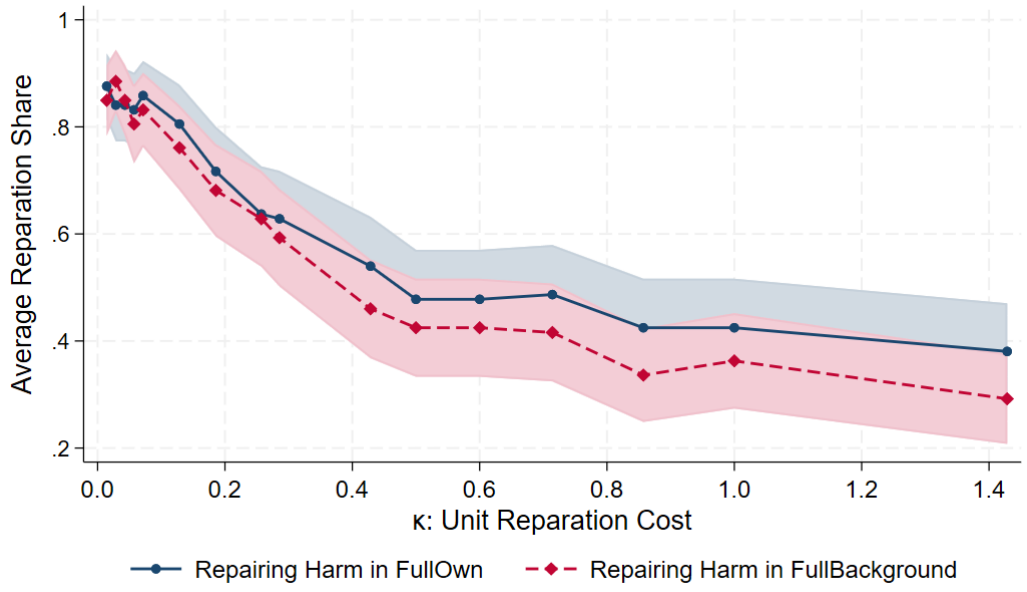


Figure 8: Fraction of Subjects Restoring Harm across Unit Reparation Cost (κ)

Notes: The figure displays the proportion of subjects choosing to restore personally caused harm (`FullOwn`, solid navy) versus existing harm (`FullBackground`, dashed cranberry) across unit reparation costs κ. Shaded bands depict 95% confidence intervals clustered at the participant level.

First, consistent with the law of demand, remediation rates in both conditions decline almost monotonically as restoring harm becomes more expensive. Second, for unit prices $\kappa \geq 0.30$, average reparation share in `FullOwn` sits systematically 4 to 9 percentage points above the reparation share in `FullBackground`. Figure 8 shows that at least some subjects pay a responsibility-premium on top of the willingness to repair harm explained by outcome-based preferences.

Note that while pointwise confidence intervals at individual price steps ($N = 99$) generate wide margins of

error, the average reparation share in `FullOwn` remains consistently above the average reparation share in `FullBackground` across all levels of κ. When pooling average reparation shares across all unit reparation costs κ the difference is highly statistically significant at the 99% level. Figure 15 in Appendix B presents the main results of the regression analysis.

Figure 9 plots the shares of subjects choosing to restore total harm, personally caused harm, background harm and restoring no harm in the reparation decision `Symmetric`. In this reparation decision, subjects caused a harm worth €7 which was identical to the value of background harm. In this decision, subjects were given a unit reparation cost equal to ~ 0.29.

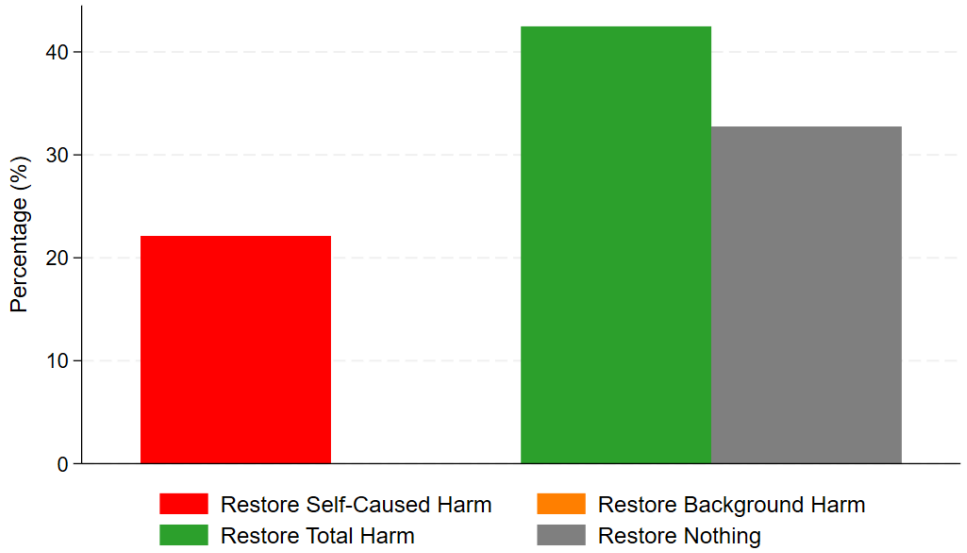


Figure 9: Choice Distribution in Reparation Stage: `Symmetric`

The large majority of subjects ($> 70\%$) repairs total harm or nothing. What is striking however, is that all partial harm restores agree on restoring personally caused harm, and no subject restores background harm of identical value. This figure provides further evidence that responsibility for negative externality drives behaviour.

5.2 Choice Patterns in `BothLow` and `BothHigh`

Figure 10 plots choice and switching patterns in `BothLow` in the left panel and `BothHigh` in the right panel. Unit reparation costs were identical under both harm generation rules. Moreover, I sorted subjects in descending order based on an average pro-sociality index that assigned the highest score to repairing total harm and the lowest score to repairing no harm. Repairing personally caused harm received the second-highest score.

Three key patterns emerge. First, a group of subjects clusters at the extremes and can be classified as

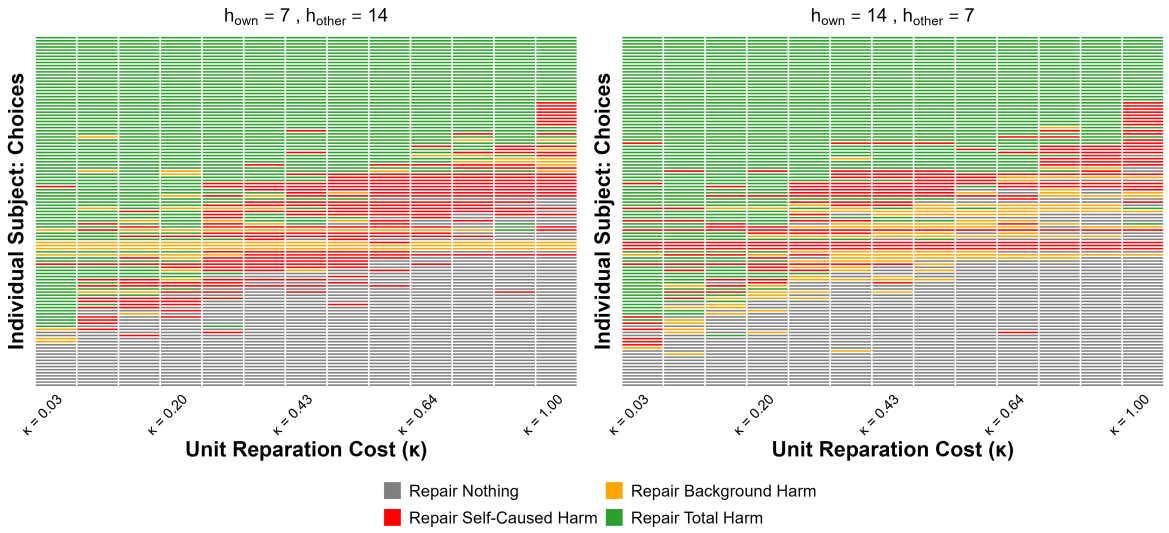


Figure 10: Individual Subject Choice Trajectories across Unit Price (κ) in CS_{both}

Notes: Each horizontal line depicts the sequence of decisions for a single participant as unit price κ increases . Panel A displays decisions under harm generation rule `BothLow` ($h_{\text{own}} = 7, h_{\text{back}} = 14$); Panel B displays decisions made under the harm generation rule `BothHigh` ($h_{\text{own}} = 14, h_{\text{back}} = 7$). Subjects are sorted vertically by their overall prosociality index (computed across both harm generation rules) and their position on the y-axis is kept identical across both panels.

either unconditional total-harm repairers or unconditional no-harm repairers. As the unit reparation cost κ increases they do never switch from repairing total harm or no harm. Second, partial harm repairers are driven by responsibility for the negative externality. As the unit reparation cost κ increases they switch from restoring total harm to restoring personally caused harm as can be seen at the center of both panels. Given the identical variation in the unit reparation costs it can be observed that some subjects in the top right corner are doubling their sacrifice to repair personally caused harm. Third, and finally, some partial harm repairers make a compromise and as the unit reparation cost doubles they switch to repair background harm. This explains the increased dominance of the yellow colour in the right panel. Yet, there is no partial-harm repairer who repairs more than the harm she has caused in the first stage.

In sum, individuals who respond to cost variations exhibit choice patterns strongly governed by a backward-looking concept of responsibility: they are systematically more willing to restore harm generated by their own action than an equivalent (or even cheaper) harm created by others.

5.3 Individual-Level Estimation of Preferences

The choice data shows at choices are partly driven by responsibility-based preferences, and thus responsibility directly influences decision-making. In this subsection, I quantify the heterogeneity in how responsibility influences decision making by structurally estimating preferences at the individual level using Maximum Likelihood and at the aggregate level estimating Finite Mixture Models.

Figure 11 plots the joint distribution of the individual-level parameters $\eta^{\text{i}}_{\text{resp}}$ and $\eta^{\text{i}}_{\text{back}}$. A single dot corre-

sponds to the estimates of $(\eta^{i}_{resp}, \eta^{i}_{back})$ of an individual subject. Subjects with identical estimates are pooled together which is captured by circles with an increasing radius around the parameter estimate.

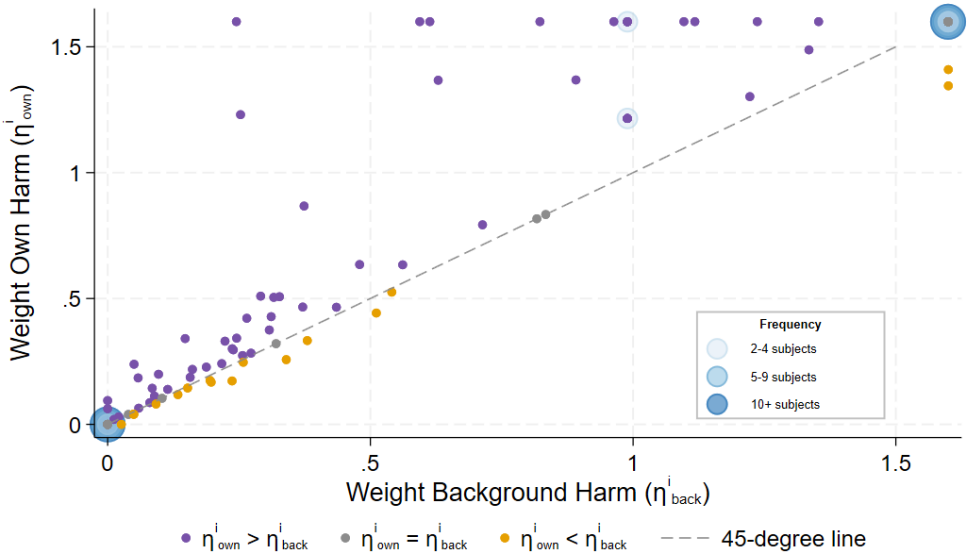


Figure 11: Individual Maximum Likelihood Estimates of the weights on unrepaired personally caused harm (η^{i}_{resp}) and unrepaired background harm (η^{i}_{back}). The 45-degree line denotes the benchmark of pure outcome-based preferences. Each dot represents one individual-level estimate. Increasing radius of the circle around the parameter estimates indicates multiple subjects with identical parameter estimates.

The 45-degree line represents the benchmark of pure outcome-based preferences ($\eta^{i}_{resp} = \eta^{i}_{back}$). If choices were entirely driven by outcome-based preferences, then subjects would distribute along this 45-degree line. However, if subjects exhibit responsibility-based preferences, then they are located above the 45-degree line. What can be observed is that responsibility directly influences reparation decisions of 75% of the sample. Some subjects exhibit a willingness to pay to repair personally caused harm that is more than three times higher than the willingness to pay to repair background harm.[20] However, there is considerable heterogeneity in preferences. First, more than 10 subjects exhaust their entire budget to repair total harm and are located at the top right corner. Second, a significant fraction of subjects can be categorized as selfish, and do not spend their earnings to repair any harm. Finally, a significant fraction of subjects is estimated very close to the 45-degree line, thus they are primarily driven by outcome-based preferences and do not account for the causal history of the negative externality. Crucially, and in line with the theory, no subject shows a considerably higher willingness to pay to repair background harm relative to personally caused harm.

Table 2 shows summary statistics for the individual-level parameter estimates of η^{i}_{resp}, η^{i}_{back}, α^{i} and μ^{i}. It shows that the median subject exhibits a willingness to pay of €0.34 to repair background harm. Instead, to repair personally caused harm, she is willing to pay €0.5 which constitutes a responsibility premium of roughly 48%. Both the mean and median of the weight on personally caused harm is higher than the mean and median of the weight on background harm while the standard deviation of both marginal distributions is very similar.

[20] For some η^{i}_{resp} is estimated at the corner of 1.5 while η^{i}_{back} is estimated to be below 0.5.

Most subjects exhibit a mix of outcome-based and responsibility-based preferences, and there is considerable heterogeneity in the strength of responsibility-based preferences. The next results show how subject's backward-looking perception of responsibility interact with background harm. Figure 12 plots the histogram of the marginal distribution of the estimates of the amplification parameter α^i. Because a large fraction of subjects unconditionally restored total harm or no harm in `BothHigh` and `BothLow`, the amplification parameter could only be identified for 68 subjects out of the studied sample.

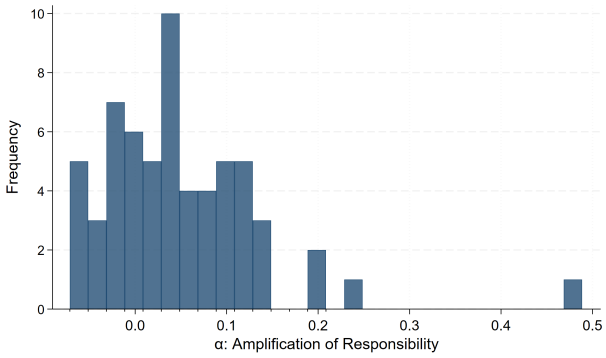


Figure 12: Marginal Distribution of Individual-Level Estimates of the Amplification Parameter α^i. Individual-level estimates are based on sample of 68 subjects who produced behavioural variation in `BothHigh` or `BothLow` that is sufficient for identification.

For the majority of subjects responsibility amplifies in the presence of background harm and for the median subject it amplifies at a rate of 4% per euro of background harm. This suggests that these subjects partially assume responsibility for the total loss of charity items. There also exist more than 10 subjects for whom responsibility attenuates in the presence of background harm. Thus, there exists heterogeneity in how responsibility interacts with background harm. The correlation between the estimated amplification parameter α^i and the estimated responsibility premium $\eta^{\text{i}}_{\text{resp}} - \eta^{\text{i}}_{\text{back}}$ is almost null, suggesting that the interaction effect does not depend on the relative strength of responsibility-based preferences. Figure 16 in Appendix B plots the correlation.

5.4 Aggregate-Level Estimation of Preferences: Results of the Finite-Mixture Model

At the aggregate level, the results of the Finite Mixture Model confirm a pervasive influence of responsibility on behaviour. Table 3 presents the parameter estimates for the representative agent model as well as for the two-type model in which I assume that the data is generated by $K = 1$ and $K = 2$ preference classes, respectively.

The representative agent exhibits a highly statistically significant responsibility premium of 29%. Her willingness to pay to restore personally caused harm sits 29% on top of her willingness to pay to restore background harm. Moreover, being estimated at 69 cents, the latter is moderately high. The responsibility of the repre-

Table 2: Summary Statistics of Estimated Utility Parameters

Parameter	**Mean**	**Median**	**Std. Dev.**	**Min**	**Max**	**N**
η_{own}: Weight self-caused harm	0.777	0.505	0.668	0.000	1.600	113
η_{back}: Weight background harm	0.654	0.340	0.619	0.000	1.600	113
μ: Choice Sensitivity	4.974	2.720	4.146	0.081	10.000	113

Table 3: Finite Mixture Model Estimations: Representative Agent ($K = 1$) and Two-Type Model ($K = 2$)

Parameter	**Representative Agent**	**2-Type Model ($K = 2$)**	
	($K = 1$)	Class 1	Class 2
Class Share (π)	1.000	0.549***	0.451***
	—	(0.041)	(0.041)
η_{own} : weight self-caused harm	0.885***	0.212***	1.500***
	(0.083)	(0.035)	(0.142)
η_{back} : weight background harm	0.685***	0.143***	1.142***
	(0.074)	(0.033)	(0.145)
α : amplification parameter	-0.071***	-0.032	0.029
	(0.002)	(0.021)	(0.123)
μ : choice sensitivity	0.165***	0.459***	0.337***
	(0.011)	(0.058)	(0.041)
p-val ($\eta_{\text{own}} = \eta_{\text{back}}$)	0.000	0.000	0.004
Log-Likelihood	-5971.60	-3611.44	
Subjects	113	113	
Observations	7910	7910	

Notes: Bootstrap standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. A z-test is used to test for $\eta^{\text{i}}_{\text{resp}} = \eta^{\text{i}}_{\text{back}}$.

sentative agent attenuates at a rate of 7% per euro of background harm.[21] Moving to the two-type model what first can be observed is that the model fit, as measured by the maximized (negative) log likelihood and the integrated completed likelihood (ICL), increases. Second, the two-type model splits subjects into a two types with almost equal shares. Both types are hybrid and exhibit responsibility-based and outcome-based preferences. Moreover, they exhibit a responsibility premium ranging between 31% and 48%. The weak hybrid type has a low baseline willingness to pay of 14 cents to repair 1 euro of background harm and the exhibited responsibility premium is 48%. Hence, she is willing to pay almost double the amount to repair personally caused harm. The strong hybrid type has a strong willingness to pay of €1.10 to repair €1 of background harm. Moreover, her willingness to pay to repair personally caused harm is estimated in the upper corner of €1.5. For both identified preference classes responsibility neither diffuses nor attenuates.[22] The three-type model splits the moderate type estimated in the two-type model into two types: a type that exhibits only responsibility-based preferences and a moderate type that exhibits both outcome-based and responsibility-based preferences. This moderate hybrid type is willing to sacrifice €0.29 to repair one euro of background harm and exhibits a 38% responsibility premium when paying to repair one euro of personally caused harm.

The results are in line with results obtained in the literature that estimate linear social preferences in dictator games or strategic interactions (Bruhin et al., 2019; Van Leeuwen and Alger, 2024).[23] First, my results confirm the key role of heterogeneity in social preferences that explains the pro-social behaviour of remediating a negative externality which also has a prominent behaviour in explaining pro-social behaviour. Second, my results also distinguish between a strong and a moderate type of social preferences. Moreover, the share of the strong hybrid is smaller than the share of the weak hybrid in the two-type model and in line with the shares of the strong altruists and moderate altruists identified in Bruhin et al. (2019) and van Leeuwen and Alger (2024). A key difference is that the strong hybrid exhibits a higher marginal willingness to remediate harm, of both sources than the identified strong altruists in Bruhin et al. (2019) and van Leeuwen and Alger (2024). While their strong altruists exhibit a marginal willingness to sacrifice to raise the payoff of another by 1€ of 0.86€-0.96€, the identified strong hybrid in my data exhibits a marginal willingness to sacrifice to restore personally caused harm of 1.5€ or possibly higher.[24]

Other studies either structurally estimate non-linear social preferences employing a CES utility function (Andreoni and Miller, 2002; Fisman et al., 2007) or they estimate a mixture-of-types model with linear social preferences (Iriberri and Rey-Biel, 2011, 2013). Irriberi and Rey-Biel (2011) estimate a marginal willingness to sacrifice of 31-33 cents for their social preference types when their payoff is higher than that of a passive other. Moreover, they separate between a selfish and a competitive type. Overall, they also identify a share of selfish and competitive types that is larger than the share of the inequity-averse and social-welfare preference type. Irriberi and Rey-Biel (2013) obtain similar results, but they do only report the distribution of preference types and not specific parameter estimates. In contrast, the share of the selfish type in Andreoni and Miller (2002) as well as Fisman et al. (2007) is smaller than that of inequity-averse or social-welfare preference types.

[21] This is an estimate at the lower corner of the allowed attenuation rate

[22] Interestingly, though, the sign is negative for the low type and positive for the high type.

[23] A difference compared to Bruhin et al. (2019) and van Leeuwen and Alger (2024) is that their structural utility specification includes behindness aversion, inequity aversion as well as competitive preferences as additional consequentialist types whereas my specification allows only for altruism as it is not the purpose of the study to discriminate among different kinds of consequentialist social preferences.

[24] The results for the weak hybrid are in line with those for the moderate altruists

Why are the strong remediators in my data more motivated to restore harm than the strong altruists in Bruhin et al. (2019) and van Leeuwen and Alger (2024)? One reason is that the context differs. In my context it is a charity who no longer distributes items to vulnerable people if subjects earn money for themselves and not another subject. Second, to make the negative externality salient to subjects, the consequences of the subject's decisions are framed as a loss to the charity and not directly as final consequences as their studies do. Third, in line with instructions from experiments that study moral behaviour in individual or market setting, subjects learn about the impact their decision has on a vulnerable person before making their decisions. This suggests that the moral aspect of their decision could be more salient to subjects when they reduce the well-being of a charity than when they reduce the well-being of an anonymous passive subject. Back-of-the-envelope calculations using experimental data studying moral behaviour in markets (Bartling et al., 2015; Kirchler et al., 2016; Bartling et al., 2019; Irlenbusch and Saxler, 2019; Falk et al., 2020; Bartling and Özdemir, 2023; Ziegler et al., 2024) provide suggestive evidence for this intuition. For example, in the single decision treatment in Irlenbusch and Saxler (2019), the average subject is indifferent between accepting a personal gain of 6.70€ instead of donating 5.20€. This corresponds to a marginal willingness to sacrifice of 1.3€ to raise the charity's payoff by 1€. Doing similar calibration exercises in the single decision-making treatments of related studies shows that the marginal willingness ranges between 0.88 (Kirchler et al., 2016) and 1.94 (Bartling and Özdemir, 2023).[25]

While an estimated finite mixture model characterizes preference heterogeneity in a parsimonious way, it has the drawback that the number of preference classes is determined exogenously, and thus in an arbitrary way. The estimated preferences at the individual level reveal considerable heterogeneity. However, to specification of the number of preference classes that are assumed to generate the choice data is an arbitrary choice made by the researcher. To minimize the scope of arbitrariness I guide the model selection by the Normalized Entropy Criterion. It penalizes ambiguous subject-to-preference-class assignment and relates it to the gain in explanatory power relative to the representative agent model. By definition, it cannot be computed for the representative agent model. Figure 13 plots the value of the Normalized Entropy Criterion as it is defined in equation (10) against the number of preference classes K specified in the representative agent model.

The NEC is minimized for $K = 2$, and thus it favours the two-type model in explaining the choice data.[26] Moreover, the NEC is almost zero for the two-type model. Since the entropy depends on the ex-post assignment probabilities, it implies unambiguous subject-to-preference-class assignment. Figure 17 in Appendix B plots the distribution of ex-post assignment probabilities in the two-type model.

[25] The single decision-making treatments in Bartling et al. (2015) and Bartling et al. (2019) provide important exceptions with an estimated marginal willingness of 0.08 and 0.22. Yet, they use a passive subject that as a third-party that is adversely affected by the decisions of active subjects and not a charity.

[26] An alternative model selection criterion is the Integrated Completed Likelihood. It also accounts for the increase in explanatory power as measured by the maximised log-likelihood and penalizes ambiguous subject-to-preference-class assignment. In addition to the NEC it also penalizes directly the inclusion of additional parameters that have to be estimated. When using the ICL as a model selection criterion, it prefers the six-type model. Given that entropy is very small even for a six-type model, it is very similar to the Bayesian Information Criterion which has been shown through Monte Carlo Simulations to perform poorly in model selection tasks (Biernacki et al., 2000; Bruhin et al., 2019) For this purpose, I do not rely on the ICL as a model selection criterion.

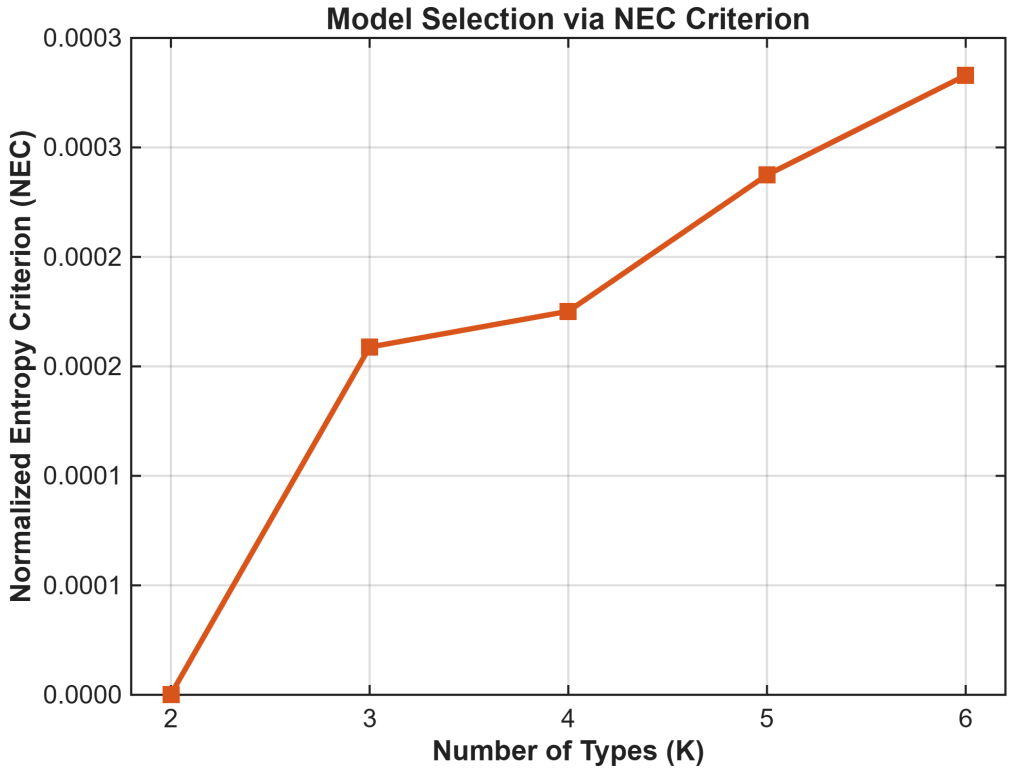


Figure 13: Normalized Entropy Criterion (NEC) in function of the number of specified preference classes in the finite mixture model.

5.5 Assessing Value Added of Allowing for Responsibility-Based Preferences

The results of the finite mixture model show a pervasive influence of responsibility on behaviour. Moreover, a two-type-model is favoured by a model selection criterion. To finally validate a model of responsibility-based preferences I will analyse the value added of allowing for such type of preferences in the structural estimation. On the one hand, it can lead to an improvement in model fit. On the other hand, it increases the number of parameters that have to be estimated, and thus model complexity. While the ICL performs poorly in model selection tasks, it is well suited to characterize the value-added of relaxing constraints on the estimated finite mixture models. I will use it in order to assess the value added of allowing for responsibility-based preferences explaining the data. For this, I first focus on a benchmark model in which I do not allow for responsibility-based preferences, i.e. $\eta^{\text{i}}_{\text{resp}} = \eta^{\text{i}}_{\text{back}}$ and $\alpha = 0$. In the second estimated model I relax the constraints and allow for the benchmark model of responsibility-based preferences, i.e. $\eta^{\text{i}}_{\text{resp}} \neq \eta^{\text{i}}_{\text{back}}$ and $\alpha = 0$. Finally, in the extended model, I also relax the constraint of α. Figure 14 plots the relative change in ICL compared to the ICL of the restricted model for the representative agent.

When allowing for $K = 1$ or $K = 2$, the ICL of the fully unrestricted model is more than 100 log points lower than that of the fully restricted model. In addition, the fully unrestricted model explains the data weakly better than the benchmark model of responsibility-based preferences. Thus, the improvement in model fit of allowing for responsibility-based preferences far outweighs the additional complexity cost resulting from estimating additional parameters, and thus responsibility-based preferences are an important driver of

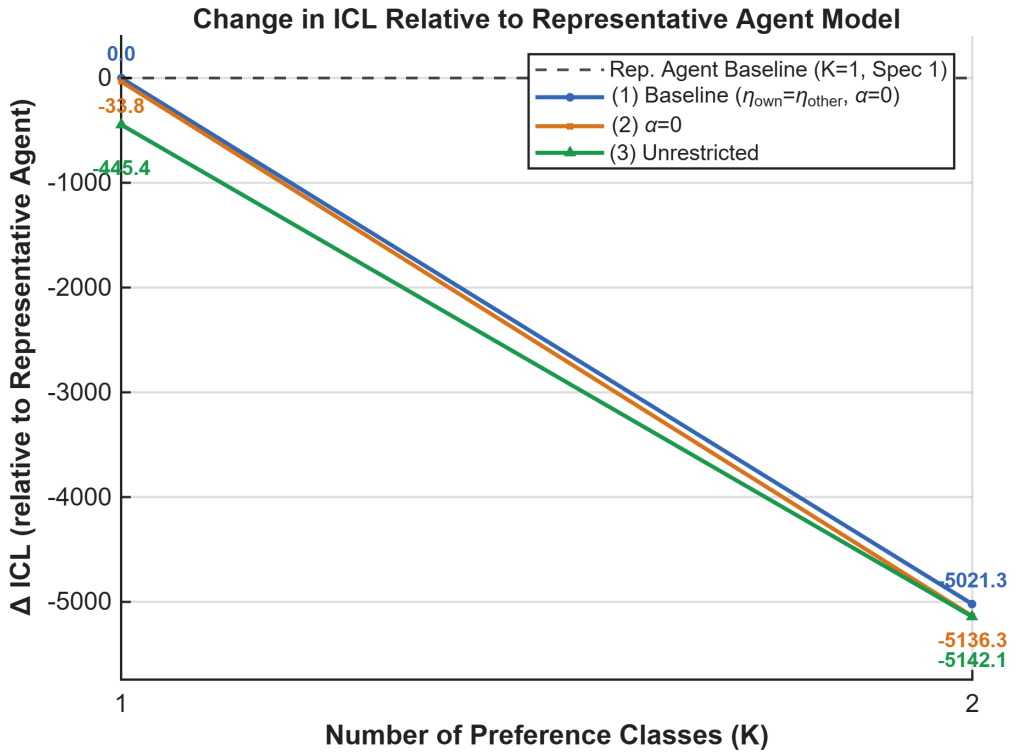


Figure 14: Comparison of the Integrated Completed Likelihood across three different estimated models: (i) a fully restricted model (restrictions: $\eta^{\text{i}}_{\text{resp}} = \eta^{\text{i}}_{\text{back}}$ and $\alpha = 0$), (ii) the benchmark model that allows for responsibility-based preferences (restriction: $\alpha = 0$), and (iii) the extended model of responsibility-based preferences (no restrictions). The models are estimated for finite mixture models using $K \in \{1, 2, 3\}$ preferences classes. The y-axis plots the relative change in ICL compared to the fully restricted model allowing for $K = 1$ preference classes.

reparation choices next to outcome-based preferences.

6 Discussion and Conclusion

In this paper, I have modelled a simple preference over the responsibility space for negative externalities. It is backward-looking in the sense that it discriminates between unrepaired personally caused harm and unrepaired background harm of unknown origin. I then showed how it can be disentangled from any outcome-based preferences in a two-stage framework in which I vary the causal history of a negative externality holding final stage-2 outcomes constant. In an extension of the theoretical framework, I allow responsibility to attenuate or – which is novel in Economics – to amplify with background harm.

Informed by the sharp revealed preference test, I designed an experiment that allows me to quantify to what extent reparation choices are driven by outcome-based and responsibility-based preferences by keeping the total harm caused to a charity constant and varying the harm generation rule. In estimating Finite Mixture Models, I find that the representative agent and both preference types identified in the two-type model exhibit a responsibility premium ranging between 29% and 48% when paying for repairing personally caused harm compared to their marginal willingness to pay for repairing background harm. At the individual level I find considerable heterogeneity in the effect of responsibility on decision-making. Some subjects want the harm caused to the charity to be repaired regardless of who has caused it. Others are heavily influenced by their own responsibility and exhibit a marginal willingness to pay to repair personally caused harm that is more than three times higher than that to repair background harm.

Why should we understand whether preferences are outcome-driven or responsibility-driven? First, it might matter for market outcomes. Suppose a seller's product pollutes the environment, but he has access to a costly offsetting technology. Does the seller want to add offsets to the products it sells? It depends whether consumers have responsibility-based or outcome-based preferences. Under the former, the seller may benefit from selling offsets to environmentally conscious consumers who otherwise suffer a moral cost from their personal footprint. Selling the offset to these consumers increases their willingness to pay for the consumption good. Crucially, however, if conscious consumers ignore harm caused by others under responsibility-based preferences the seller does not want to sell offsets to selfish consumers. Hence, the externality caused by the products sold to selfish consumers remains unaddressed in these markets. Under outcome-based preferences, the seller could instead leverage the conscious consumers' preferences to mitigate a negative externality regardless of who has caused it. In other words, responsibility-based preferences limit the scope of corporate philanthropy even though it might be efficient (Bénabou and Tirole, 2010).[27]

Second, the sharp revealed preference test can be directly applied to the context of voluntary carbon offsets to study its demand, especially to what extent it is driven by efficiency or equity concerns as opposed to concerns related to personal responsibility. This matters because the market is growing rapidly. It has doubled its size over roughly the last half decade and is projected to grow to a market volume of $50 billion by 2030 (Blaufelder et al., 2021). It can be flexibly adapted to other harm generation rules in which there exists uncertainty about the individual impact but there exist focal points that allow to anchor own perceived responsibility such as the average carbon emissions of the consumption good.

[27] I formalize these intuitions in an IO framework (Guias, 2026).

One key contribution of this paper is to formalize responsibility-based preferences. The proposed mechanism through which responsibility directly influences decision making captures the core distinction between final outcomes for which a decision maker perceives herself as responsible and outcomes for which she does not. Once the core mechanism has been theoretically introduced and experimentally validated it can be extended. A simple extension is to allow for the interaction of responsibility-based preferences with social image concerns. How does responsibility shape decision-making under observability? This may depend on the decision maker's beliefs about the other's personal norms, and to what extent the other considers the assumption of responsibility for the harm caused by others to be praiseworthy. This mechanism is relevant for understanding corporate social responsibility practices. Whether firms seek to maximize their social impact or minimize own responsibility depends to what extent stakeholders stigmatize the generation of negative externalities and praise practices that maximize social impact. Another interesting research avenue is how responsibility-based preferences interact with social identity. To what extent does an individual perceive herself as responsible for harm caused by in-group members and how does it shape decision-making? This interaction effect plays an important role in understanding how societies vote for reparations of harm caused by previous generations.

While this paper has proposed a core mechanism for the direct influence of responsibility on behaviour, it provides by no means a general theory of responsibility-based preferences. First, it focuses on a setting of negative externality, and hence responsibility is framed in terms of losses relative to a reference point. Therefore, it remains silent about how perceived responsibility for gains influences decision making. Moreover, it focuses on a linear responsibility-based preference and does not allow for non-linear preferences. Regarding the experiment, the identification strategy relies on subjects being willing to accept a monetary benefit that causes harm to the charity and roughly 30% of the sample decided to not accept the monetary benefit. While the sample studied in the experiment is relevant for market outcomes and policy making, the experimental design does not allow for making claims about preferences at the population level. Research on the methodological side may improve the experimental design to allow for generalizations at the population level.

In sum, this paper provides granular evidence that responsibility for negative externality influences decision-making directly. On the one hand, this increases individual willingness to pay to mitigate personally caused harm. On the other hand, it limits the moral domain to personally caused harm, and therefore harm that is not personally caused is ignored under these preferences. There exists scope for research to understand how responsibility directly shapes decision-making.

References

Akerlof, G. and Kranton, R. (2000). Economics and Identity. *The Quarterly Journal of Economics*, 115(3).

Alger, I. and Rivero-Wildemauwe, J. I. (2024). Doing the right thing (or not) in a lemons-like situation: on the role of social preferences and Kantian moral concerns. *Working Paper*.

Alger, I. and Weibull, J. (2013). Homo Moralis–Preference Evolution Under Incomplete Information and Assortative Matching. *Econometrica*, 81(6):2269–2302.

Almas, I., Cappelen, A. W., and Tungodden, B. (2020). Cutthroat Capitalism versus Cuddly Socialism: Are Americans More Meritocratic and Efficiency-Seeking than Scandinavians? *The Journal of Political Economy*, 128(5):1753–1788.

Andre, P. (2025). Shallow Meritocracy. *Review of Economic Studies*, 92(2):772–807.

Andreoni, J. (1990). Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving. *The Economic Journal*, 100(401):464.

Andreoni, J. and Miller, J. (2002). Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism. *Econometrica*, 70(2):737–753.

Bartling, B. and Fischbacher, U. (2012). Shifting the Blame: On Delegation and Responsibility. *The Review of Economic Studies*, 79(1):67–87.

Bartling, B., Valero, V., and Weber, R. (2019). On the scope of externalities in experimental markets. *Experimental Economics*, 22(3):610–624.

Bartling, B., Weber, R. A., and Yao, L. (2015). Do Markets Erode Social Responsibility? *The Quarterly Journal of Economics*, 130(1):219–266.

Bartling, B. and Özdemir, Y. (2023). The limits to moral erosion in markets: Social norms and the replacement excuse. *Games and Economic Behavior*, 138:143–160.

Battigalli, P. and Dufwenberg, M. (2007). Guilt in Games. *American Economic Review*, 97(2).

Becker, G. S. (1974). A Theory of Social Interactions. *The Journal of Political Economy*, 82(6):1063–1093.

Benabou, R. and Tirole, J. (2006). Incentives and Prosocial Behavior. *The American Economic Review*, 96(5):1652–1678.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725.

Blaufelder, C., Levy, C., Mannion, P., and Pinner, D. (2021). A blueprint for scaling voluntary carbon markets to meet the climate challenge.

Bolton, G. E. and Ockenfels, A. (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *The American Economic Review*, 90(1):166–193.

Brekke, K. A., Kverndokk, S., and Nyborg, K. (2003). An economic model of moral motivation. *Journal of Public Economics*, 87(9-10):1967–1983.

Bruhin, A., Fehr, E., and Schunk, D. (2019). The many Faces of Human Sociality: Uncovering the Distribution and Stability of Social Preferences. *Journal of the European Economic Association*, 17(4):1025–1069.

Bénabou, R., Falk, A., Henkel, L., and Tirole, J. (2024). Eliciting Moral Preferences under Image Concerns: Theory and Experiment. *Working Paper*.

Bénabou, R. and Henkel, L. (2025). Identity as Self-Image.

Bénabou, R. and Tirole, J. (2010). Individual and Corporate Social Responsibility. *Economica*, 77(305):1–19.

Calel, R., Colmer, J., Dechezleprêtre, A., and Glachant, M. (2025). Do Carbon Offsets Offset Carbon? *American Economic Journal: Applied Economics*, 17(1):1–40.

Cappelen, A. W., Hole, A. D., Sørensen, E. ., and Tungodden, B. (2007). The Pluralism of Fairness Ideals: An Experimental Approach. *The American Economic Review*, 97(3):818–827.

Cappelen, A. W., Konow, J., Sørensen, E. ., and Tungodden, B. (2013a). Just Luck: An Experimental Study of Risk-Taking and Fairness. *American Economic Review*, 103(4):1398–1413.

Cappelen, A. W., Moene, K. O., Sørensen, E. ., and Tungodden, B. (2013b). Needs versus Entitlements-An International Fairness Experiment. *Journal of the European Economic Association*, 11(3):574–598.

Cappelen, A. W., Sørensen, E. ., and Tungodden, B. (2010). Responsibility for what? Fairness and individual responsibility. *European Economic Review*, 54(3):429–441.

Charness, G. and Rabin, M. (2002). Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, 117(3):817–869.

Cicero, M. T. (1913). *De Oficiis (On Duty)*. Cambridge, MA, harvard university press edition.

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80.

De Groot, J. I. M. and Steg, L. (2009). Morality and Prosocial Behavior: The Role of Awareness, Responsibility, and Norms in the Norm Activation Model. *The Journal of Social Psychology*, 149(4):425–449.

Dillenberger, D. and Sadowski, P. (2012). Ashamed to be selfish. *Theoretical Economics*, 7(1):99–124.

Ecosystem Market Place (2025). State of the Voluntary Market 2025.

Ellingsen, T. and Mohlin, E. (2026). A Model of Social Duties. *Journal of Political Economy*, 134(5):1307–1346.

Falk, A., Neuber, T., and Szech, N. (2020). Diffusion of Being Pivotal and Immoral Outcomes. *The Review of Economic Studies*, 87(5):2205–2229.

Falk, A. and Szech, N. (2013). Morals and Markets. *Science*, 340.

Fehr, E. and Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.

Fisman, R., Kariv, S., and Markovits, D. (2007). Individual Preferences for Giving. *The American Economic Review*, 97(5):1858–1876.

Gneezy, U., Imas, A., and Madarász, K. (2014). Conscience Accounting: Emotion Dynamics and Social Behavior. *Management Science*, 60(11):2645–2658.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., and Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, 111(3):364–371.

Hainmueller, J., Hiscox, M. J., and Sequeira, S. (2015). Consumer Demand for Fair Trade: Evidence from a Multistore Field Experiment. *Review of Economics and Statistics*, 97(2):242–256.

Halpern, J. Y. (2016). *Actual causality*. The MIT Press, Cambridge, Massachusetts.

Hamman, J. R., Loewenstein, G., and Weber, R. A. (2010). Self-Interest through Delegation: An Additional Rationale for the Principal-Agent Relationship. *American Economic Review*, 100(4):1826–1846.

Hart, O. and Zingales, L. (2017). Companies Should Maximize Shareholder Welfare Not Market Value. *Journal of Law, Finance and Accounting*, 2(2):247–275.

Iriberri, N. and Rey-Biel, P. (2011). The role of role uncertainty in modified dictator games. *Experimental Economics*, 14(2):160–180.

Iriberri, N. and Rey-Biel, P. (2013). Elicited beliefs and social information in modified dictator games: What do dictators believe other dictators do?: Beliefs and information in dictator games. *Quantitative Economics*, 4(3):515–547.

Irlenbusch, B. and Saxler, D. J. (2019). The role of social information, market framing, and diffusion of responsibility as determinants of socially responsible behavior. *Journal of Behavioral and Experimental Economics*, 80:141–161.

Jagelka, T. (2024). Are Economists' Preferences Psychologists' Personality Traits? A Structural Approach. *Journal of Political Economy*, 132(3):910–970.

Jakob, M., Kübler, D., Steckel, J. C., and Van Veldhuizen, R. (2017). Clean up your own mess: An experimental study of moral responsibility and efficiency. *Journal of Public Economics*, 155:138–146.

Kirchler, M., Huber, J., Stefan, M., and Sutter, M. (2016). Market Design and Moral Behavior. *Management Science*, 62(9):2615–2625.

Noor, J. and Ren, L. (2023). Temptation and guilt. *Games and Economic Behavior*, 140:272–295.

Pigors, M. and Rockenbach, B. (2016). Consumer Social Responsibility. *Management Science*, 62(11):3123–3137.

Rodemeier, M. (2026). Willingness to Pay for Carbon Mitigation: Field Evidence from the Market for Carbon Offsets. *The Review of Financial Studies*, 39(1):1–29.

Schneider, F., Brun, F., and Weber, R. A. (2024). Sorting and Wage Premiums in Immoral Work. *The Review of Economics and Statistics*, pages 1–46.

Schulze-Tilling, A. (2025). The Effectiveness of Carbon Labels.

Schwartz, S. H. (1977). Normative Influences on Altruism. In *Advances in Experimental Social Psychology*, volume 10, pages 221–279. Elsevier.

Semken, C. (2026). The Marginal Impact of Emission Reductions: Estimates, Beliefs and Behavior.

Smith, A. (1759). *The Theory of Moral Sentiments*. Prometheus Books. Amherst, NY.

Sutter, M., Huber, J., Kirchler, M., Stefan, M., and Walzl, M. (2020). Where to look for the morals in markets? *Experimental Economics*, 23(1):30–52.

Talbert, M. (2019). Moral Responsibility.

Van Leeuwen, B. and Alger, I. (2024). Estimating Social Preferences and Kantian Morality in Strategic Interactions. *Journal of Political Economy Microeconomics*, 2(4):665–706.

Ziegler, A. G. B., Romagnoli, G., and Offerman, T. (2024). Morals in Multi-Unit Markets. *Journal of the European Economic Association*, 22(5):2225–2260.

A Appendix

A.1 Appendix: Proof of Proposition 1

Proof of Proposition 1. The decision maker chooses a reparation level $c \in [0, H]$, where total historical harm is $H = bh_{\text{own}} + h_{\text{back}}$. Her utility function is defined as:

$$u(c; b, h_{\text{own}}, h_{\text{back}}) = x_{\text{DM}}^{\text{RS}}(b, c) + \eta \cdot x_r^{\text{RS}}(b, c) - \eta_{\text{resp}} \cdot \rho(b, c; h_{\text{own}})$$

Substituting stage-2 payoffs $x_{\text{DM}}^{\text{RS}}(b, c) = by - \kappa c$, $x_r^{\text{RS}}(b, c) = \omega_r - h_{\text{back}} - bh_{\text{own}} + c$, and the backward-looking responsibility function $\rho(b, c; h_{\text{own}}) = \max\{bh_{\text{own}} - c, 0\}$, utility simplifies to:

$$u(c) = C_0 + (\eta - \kappa)c - \eta_{\text{resp}} \max\{bh_{\text{own}} - c, 0\}$$

where $C_0 \equiv by + \eta(\omega_r - h_{\text{back}} - bh_{\text{own}})$ is independent of c.
The function $u(c)$ is continuous on $[0, H]$ and non-differentiable at $c = bh_{\text{own}}$. We evaluate the derivative piecewise:

1. **Case 1:** $c \in [0, bh_{\text{own}}]$**:** We have $\rho(b, c; h_{\text{own}}) = bh_{\text{own}} - c$, giving:

$$u_1(c) = C_0 - \eta_{\text{resp}} bh_{\text{own}} + (\eta + \eta_{\text{resp}} - \kappa)c \implies \frac{\partial u_1}{\partial c} = \eta + \eta_{\text{resp}} - \kappa$$

2. **Case 2:** $c \in (bh_{\text{own}}, H]$**:** Here $\rho(b, c; h_{\text{own}}) = 0$, giving:

$$u_2(c) = C_0 + (\eta - \kappa)c \implies \frac{\partial u_2}{\partial c} = \eta - \kappa$$

Evaluating the sign of marginal utility across parameter spaces of unit reparation cost κ:

- **Part (a) Purely Outcome-Driven Reparation ($\kappa \leq \eta$):** Since $\kappa \leq \eta$, $\frac{\partial u_1}{\partial c} \geq \eta_{\text{resp}} \geq 0$ and $\frac{\partial u_2}{\partial c} \geq 0$. Utility $u(c)$ is non-decreasing on $[0, H]$ (and strictly increasing when $\kappa < \eta$). Thus, $u(c)$ is maximized at $c^* = H = bh_{\text{own}} + h_{\text{back}}$.
- **Part (b) Purely Responsibility-Driven Reparation ($\eta < \kappa \leq \eta + \eta_{\text{resp}}$):** Here $\frac{\partial u_1}{\partial c} \geq 0$ on $[0, bh_{\text{own}}]$ and $\frac{\partial u_2}{\partial c} < 0$ on $(bh_{\text{own}}, H]$. Utility increases up to $c = bh_{\text{own}}$ and strictly decreases thereafter. The global maximum is uniquely attained at $c^* = bh_{\text{own}}$.
- **Part (c) No Reparation ($\kappa > \eta + \eta_{\text{resp}}$):** Here $\frac{\partial u_1}{\partial c} < 0$ and $\frac{\partial u_2}{\partial c} < 0$ globally across $[0, H]$. Utility $u(c)$ is strictly decreasing in c. The global maximum is uniquely attained at $c^* = 0$.

□

A.2 Proof of Proposition 2: Sharp Revealed Preference Test

Proof of Proposition 2. Fix a total negative externality $\bar{H} = h_{\text{own}} + h_{\text{back}} > 0$, and suppose the decision maker (DM) accepts the monetary benefit in Stage 1 ($b = 1$). Let $\gamma \in [0, 1]$ denote the DM's relative causal share, such that personally caused harm is $h_{\text{own}} = \gamma \bar{H}$ and background harm is $h_{\text{back}} = (1 - \gamma)\bar{H}$.

Part 1: Invariance of Consequence-Based Utility Functions.

Let $U(x_{\text{DM}}, x_r)$ be any strictly outcome-based (consequentialist) utility function defined solely over final Stage-2 payoffs. Conditional on $b = 1$, the final payoffs to the DM and the passive recipient in the Reparation Stage given any reparation choice $c \in [0, \bar{H}]$ are:

$$x_{\text{DM}}^{\text{RS}}(1, c) = y - \kappa c,$$
$$x_r^{\text{RS}}(1, c) = \omega_r - \bar{H} + c.$$

Note that for any fixed initial endowment ω_r, monetary benefit y, unit price κ, and total harm $\bar{H}$, the payoff pair $(x_{\text{DM}}^{\text{RS}}(1, c), x_r^{\text{RS}}(1, c))$ depends strictly on c and is completely independent of the causal composition parameter γ.
Consequently, the utility maximization problem for a purely outcome-based decision maker is:

$$c^*(\kappa, \gamma) \in \arg\max_{c \in [0, \bar{H}]} U\left(y - \kappa c,\, \omega_r - \bar{H} + c\right).$$

Because neither the objective function $U(\cdot, \cdot)$ nor the feasible choice domain $[0, \bar{H}]$ contains γ, the set of utility-maximizing choices $\arg\max U$ is invariant to variations in γ. Assuming any deterministic choice selection rule over the outcome space in the event of indifference, the selected choice c^* is invariant to γ. Differentiating c^* with respect to γ yields:

$$\frac{\partial c^*}{\partial \gamma} = 0 \quad \text{for all } \gamma \in [0, 1].$$

Thus, no strictly outcome-based utility function $U(x_{\text{DM}}, x_r)$ can rationalize $\frac{\partial c^*}{\partial \gamma} > 0$.

Part 2: Comparative Statics under Responsibility-Based Preferences.
Now suppose preferences are given by Equation (2), incorporating both outcome-based concerns and backward-looking responsibility:

$$u(c; b = 1, h_{\text{own}}, h_{\text{back}}) = x_{\text{DM}}^{\text{RS}}(1, c) + \eta_{\text{outcome}} \cdot x_r^{\text{RS}}(1, c) - \eta_{\text{resp}} \cdot \rho(1, c; h_{\text{own}}).$$

Substituting $h_{\text{own}} = \gamma\bar{H}$, $h_{\text{back}} = (1-\gamma)\bar{H}$, and $\rho(1, c; \gamma\bar{H}) = \max\{\gamma\bar{H} - c, 0\}$, the utility function simplifies to:

$$u(c) = y - \kappa c + \eta_{\text{outcome}}(\omega_r - \bar{H} + c) - \eta_{\text{resp}} \max\{\gamma\bar{H} - c, 0\}.$$

Applying the piecewise optimization results from Proposition 1, the optimal reparation choice $c^*(\kappa, \gamma)$ depends on the unit reparation cost κ:

1. **Outcome-Based Region ($\kappa \leq \eta_{\text{outcome}}$):** The DM sets $c^* = \bar{H}$. Since total harm $\bar{H}$ is fixed with respect to γ, we have $\frac{\partial c^*}{\partial \gamma} = 0$.

2. **Responsibility-Based Region ($\eta_{\text{outcome}} < \kappa \leq \eta_{\text{outcome}} + \eta_{\text{resp}}$):** The DM repairs exclusively personally caused harm, setting $c^* = h_{\text{own}} = \gamma\bar{H}$. Differentiating c^* with respect to γ yields:

$$\frac{\partial c^*}{\partial \gamma} = \frac{\partial(\gamma\bar{H})}{\partial \gamma} = \bar{H} > 0.$$

3. **Selfish Region ($\kappa > \eta_{\text{outcome}} + \eta_{\text{resp}}$):** The DM sets $c^* = 0$, yielding $\frac{\partial c^*}{\partial \gamma} = 0$.

Thus, when $\kappa \in (\eta_{\text{outcome}}, \eta_{\text{outcome}} + \eta_{\text{resp}}]$, the DM responds to an increase in her relative causal share γ by increasing reparation at a rate equal to total harm $\bar{H}$ ($\frac{\partial c^*}{\partial \gamma} = \bar{H} > 0$), specifically tracking and repairing personally caused harm ($c^* = h_{\text{own}}$). Otherwise, $\frac{\partial c^*}{\partial \gamma} = 0$. □

A.3 Extended Framework: Proof of Proposition 3 and Sharp Revealed Preference Test

Proof of Proposition 3. The decision maker chooses a reparation level $c \in [0, H]$, where total historical harm is $H = bh_{\text{own}} + h_{\text{back}}$. Under the extended model with responsibility attenuation/amplification, her utility function is defined as:

$$u(c; b, h_{\text{own}}, h_{\text{back}}) = x_{\text{DM}}^{\text{RS}}(b, c) + \eta_{\text{outcome}} \cdot x_r^{\text{RS}}(b, c) - \eta_{\text{resp}} \cdot \delta(h_{\text{back}}; \alpha) \cdot \rho(b, c; h_{\text{own}})$$

where $\delta(h_{\text{back}}; \alpha) = \exp(\alpha h_{\text{back}})$ is the attenuation/amplification function. Substituting stage-2 payoffs $x_{\text{DM}}^{\text{RS}}(b, c) = by - \kappa c$, $x_r^{\text{RS}}(b, c) = \omega_r - h_{\text{back}} - bh_{\text{own}} + c$, and the backward-looking responsibility function $\rho(b, c; h_{\text{own}}) = \max\{bh_{\text{own}} - c, 0\}$, utility simplifies to:

$$u(c) = C_0 + (\eta_{\text{outcome}} - \kappa)c - \eta_{\text{resp}}\delta(h_{\text{back}}; \alpha) \max\{bh_{\text{own}} - c, 0\}$$

where $C_0 \equiv by + \eta_{\text{outcome}}(\omega_r - h_{\text{back}} - bh_{\text{own}})$ is independent of c.
Because $\exp(\alpha h_{\text{back}}) > 0$ for all $\alpha \in \mathbb{R}$ and all $h_{\text{back}} \geq 0$, the composite multiplier $\eta_{\text{resp}}\delta(h_{\text{back}}; \alpha)$ is strictly positive. The function $u(c)$ is continuous on $[0, H]$ and non-differentiable at $c = bh_{\text{own}}$. We evaluate the derivative piecewise:

1. **Case 1:** $c \in [0, \mathbf{bh_{own}}]$**:** Here $\rho(b, c; h_{\text{own}}) = bh_{\text{own}} - c$, giving:

$$u_1(c) = C_0 - \eta_{\text{resp}}\delta(h_{\text{back}}; \alpha)bh_{\text{own}} + (\eta_{\text{outcome}} + \eta_{\text{resp}}\delta(h_{\text{back}}; \alpha) - \kappa)\, c \implies \frac{\partial u_1}{\partial c} = \eta_{\text{outcome}} + \eta_{\text{resp}}\delta(h_{\text{back}}; \alpha) - \kappa$$

2. **Case 2:** $c \in (\mathbf{bh_{own}}, H]$**:** Here $\rho(b, c; h_{\text{own}}) = 0$, giving:

$$u_2(c) = C_0 + (\eta_{\text{outcome}} - \kappa)c \implies \frac{\partial u_2}{\partial c} = \eta_{\text{outcome}} - \kappa$$

Evaluating the sign of marginal utility across parameter spaces of unit reparation cost κ:

- **Part (a1) Purely Outcome-Driven Reparation ($\kappa \leq \eta_{\text{outcome}}$):** Since $\kappa \leq \eta_{\text{outcome}}$, we have $\frac{\partial u_1}{\partial c} \geq \eta_{\text{resp}}\delta(h_{\text{back}}; \alpha) > 0$ and $\frac{\partial u_2}{\partial c} \geq 0$. Utility $u(c)$ is non-decreasing on $[0, H]$ (and strictly increasing when $\kappa < \eta_{\text{outcome}}$). Thus, $u(c)$ is maximized at $c^* = H = bh_{\text{own}} + h_{\text{back}}$.
- **Part (a2) Purely Responsibility-Driven Reparation ($\eta_{\text{outcome}} < \kappa \leq \eta_{\text{outcome}} + \delta(h_{\text{back}}; \alpha)\eta_{\text{resp}}$):** Here $\frac{\partial u_1}{\partial c} \geq 0$ on $[0, bh_{\text{own}}]$ and $\frac{\partial u_2}{\partial c} < 0$ on $(bh_{\text{own}}, H]$. Utility increases up to $c = bh_{\text{own}}$ and strictly decreases thereafter. The global maximum is uniquely attained at $c^* = bh_{\text{own}}$.
- **Part (a3) No Reparation ($\kappa > \eta_{\text{outcome}} + \delta(h_{\text{back}}; \alpha)\eta_{\text{resp}}$):** Here $\frac{\partial u_1}{\partial c} < 0$ and $\frac{\partial u_2}{\partial c} < 0$ globally across $[0, H]$. Utility $u(c)$ is strictly decreasing in c. The global maximum is uniquely attained at $c^* = 0$.

□

Proposition 4. *Let $x_{DM} = x_{DM}^{RS}(b,c)$ and $x_r = x_r^{RS}(b,c)$. Suppose the DM is unobserved by others, observes exogenous background harm of unknown origin (h_{back}) and cannot plan reparation choices in the first stage. Moreover, fix a total negative externality $\bar{H} = h_{own} + h_{back}$ and suppose the DM takes the monetary benefit in stage 1. Finally, let $\gamma^* = 1 - \frac{1}{\alpha \bar{H}} \ln\left(\frac{\kappa - \eta}{\eta_{resp}^1}\right)$. Then the following is true in the second stage:*

1) *No strictly consequence-based utility function, $U(x_{DM}, x_r)$, can rationalize $\frac{\partial c^*}{\partial \gamma} > 0$.*

2) *Under the presence of both outcome-based and responsibility-based preferences with diffusion and amplification of perceived responsibility the following holds:*

- *Under **Responsibility amplification** ($\alpha > 0$):*
  - *If $\eta_{outcome} < \kappa \leq \eta_{outcome} + \eta_{resp}$, then $c^* = h_{own}$ and $\frac{\partial c^*}{\partial \gamma}\Big|_H = H$ globally for all $\gamma \in [0,1]$.*
  - *If $\eta_{outcome} + \eta_{resp} < \kappa \leq \eta_{outcome} + \eta_{resp} e^{\alpha H}$, then $c^* = h_{own}$ and $\frac{\partial c^*}{\partial \gamma}\Big|_H = H$ for $\gamma < \gamma^*$; otherwise $c^* = 0$.*
- *Under **Responsibility attenuation** ($\alpha < 0$):*
  - *If $\eta_{outcome} < \kappa \leq \eta_{outcome} + \eta_{resp} e^{\alpha H}$, then $c^* = h_{own}$ and $\frac{\partial c^*}{\partial \gamma}\Big|_H = H$ globally for all $\gamma \in [0,1]$.*
  - *If $\eta_{outcome} + \eta_{resp} e^{\alpha H} < \kappa \leq \eta_{outcome} + \eta_{resp}$, then $c^* = h_{own}$ and $\frac{\partial c^*}{\partial \gamma}\Big|_H = H$ for $\gamma > \gamma^*$; otherwise $c^* = 0$.*

Proof of Proposition 4. Fix a total negative externality $\bar{H} = h_{\text{own}} + h_{\text{back}} > 0$, and suppose the decision maker (DM) accepts the monetary benefit in Stage 1 ($b = 1$). Let $\gamma \in [0,1]$ denote the DM's relative causal share, such that personally caused harm is $h_{\text{own}} = \gamma \bar{H}$ and background harm is $h_{\text{back}} = (1-\gamma)\bar{H}$.

Part 1: See proof for Proposition 2.

Part 2: Comparative Statics under Extended Responsibility-Based Preferences. Now suppose preferences incorporate both outcome-based concerns and backward-looking responsibility with attenuation or amplification as given by Equation (3):

$$u(c; b = 1, h_{\text{own}}, h_{\text{back}}) = x_{\text{DM}}^{\text{RS}}(1,c) + \eta_{\text{outcome}} \cdot x_r^{\text{RS}}(1,c) - \eta_{\text{resp}} \cdot \delta(h_{\text{back}}; \alpha) \cdot \rho(1, c; h_{\text{own}}),$$

where $\delta(h_{\text{back}}; \alpha) = \exp(\alpha h_{\text{back}})$. Substituting $h_{\text{own}} = \gamma \bar{H}$ and $h_{\text{back}} = (1-\gamma)\bar{H}$, utility simplifies to:

$$u(c) = y - \kappa c + \eta_{\text{outcome}}(\omega_r - \bar{H} + c) - \eta_{\text{resp}} \exp\left(\alpha(1-\gamma)\bar{H}\right) \max\{\gamma \bar{H} - c, 0\}.$$

From Proposition 3, for $\kappa > \eta_{\text{outcome}}$, the DM repairs personally caused harm $c^* = h_{\text{own}} = \gamma \bar{H}$ if and only if the unit reparation cost satisfies:

$$\kappa \leq \eta_{\text{outcome}} + \eta_{\text{resp}} \exp\left(\alpha(1-\gamma)\bar{H}\right).$$

Rearranging this condition yields:

$$\ln\left(\frac{\kappa - \eta_{\text{outcome}}}{\eta_{\text{resp}}}\right) \leq \alpha(1-\gamma)\bar{H}.$$

We analyze the optimal choice c^* and derivative $\frac{\partial c^*}{\partial \gamma}\Big|_{\bar{H}}$ across cases of α:

1. **Responsibility Amplification ($\alpha > 0$):** Dividing by $\alpha\bar{H} > 0$ yields:

$$\gamma \leq 1 - \frac{1}{\alpha\bar{H}} \ln\left(\frac{\kappa - \eta_{\text{outcome}}}{\eta_{\text{resp}}}\right) \equiv \gamma^*.$$

   - If $\eta_{\text{outcome}} < \kappa \leq \eta_{\text{outcome}} + \eta_{\text{resp}}$, then $\ln\left(\frac{\kappa - \eta_{\text{outcome}}}{\eta_{\text{resp}}}\right) \leq 0$, which implies $\gamma^* \geq 1$. Hence, $\gamma \leq \gamma^*$ holds globally for all $\gamma \in [0, 1]$. The DM sets $c^* = h_{\text{own}} = \gamma\bar{H}$ and $\frac{\partial c^*}{\partial \gamma}\Big|_{\bar{H}} = \bar{H} > 0$.
   - If $\eta_{\text{outcome}} + \eta_{\text{resp}} < \kappa \leq \eta_{\text{outcome}} + \eta_{\text{resp}} e^{\alpha\bar{H}}$, then $0 \leq \gamma^* < 1$. For $\gamma < \gamma^*$, the DM repairs personally caused harm ($c^* = h_{\text{own}} = \gamma\bar{H}$ and $\frac{\partial c^*}{\partial \gamma}\Big|_{\bar{H}} = \bar{H}$); for $\gamma \geq \gamma^*$, she chooses no reparation ($c^* = 0$ and $\frac{\partial c^*}{\partial \gamma}\Big|_{\bar{H}} = 0$).

2. **Responsibility Attenuation ($\alpha < 0$):** Dividing by $\alpha\bar{H} < 0$ flips the inequality:

$$\gamma \geq 1 - \frac{1}{\alpha\bar{H}} \ln\left(\frac{\kappa - \eta_{\text{outcome}}}{\eta_{\text{resp}}}\right) \equiv \gamma^*.$$

   - If $\eta_{\text{outcome}} < \kappa \leq \eta_{\text{outcome}} + \eta_{\text{resp}} e^{\alpha\bar{H}}$, then $\ln\left(\frac{\kappa - \eta_{\text{outcome}}}{\eta_{\text{resp}}}\right) \leq \alpha\bar{H} < 0$, which implies $\gamma^* \leq 0$. Thus, $\gamma \geq \gamma^*$ holds globally for all $\gamma \in [0, 1]$, yielding $c^* = h_{\text{own}} = \gamma\bar{H}$ and $\frac{\partial c^*}{\partial \gamma}\Big|_{\bar{H}} = \bar{H} > 0$.
   - If $\eta_{\text{outcome}} + \eta_{\text{resp}} e^{\alpha\bar{H}} < \kappa \leq \eta_{\text{outcome}} + \eta_{\text{resp}}$, then $0 \leq \gamma^* < 1$. For $\gamma > \gamma^*$, the DM repairs personally caused harm ($c^* = h_{\text{own}} = \gamma\bar{H}$ and $\frac{\partial c^*}{\partial \gamma}\Big|_{\bar{H}} = \bar{H}$); for $\gamma \leq \gamma^*$, she chooses no reparation ($c^* = 0$ and $\frac{\partial c^*}{\partial \gamma}\Big|_{\bar{H}} = 0$).

□

B Appendix: Experimental Results

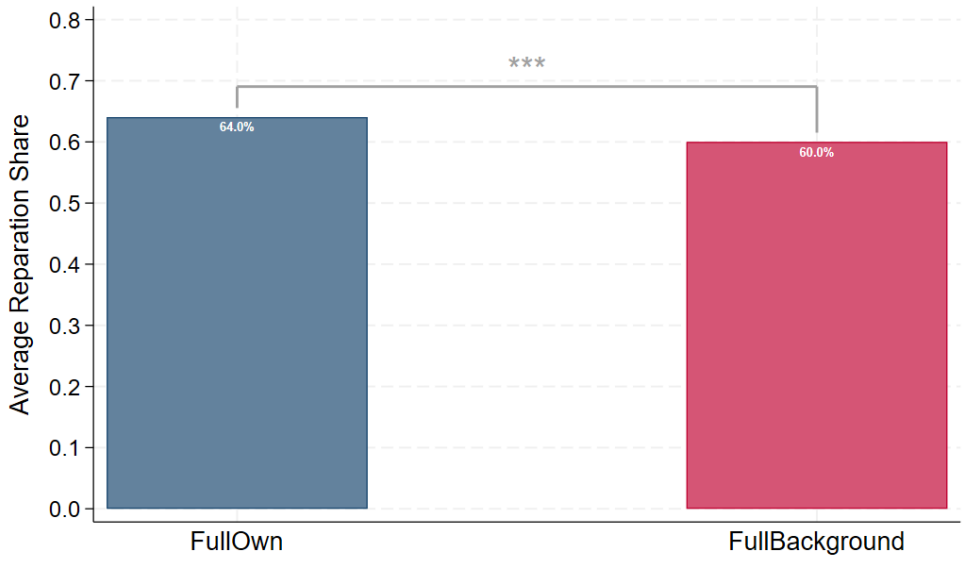


Figure 15: Regression Plot: Average Reparation Shares in `FullOwn` and `FullBackground` pooled across all unit reparation costs κ.

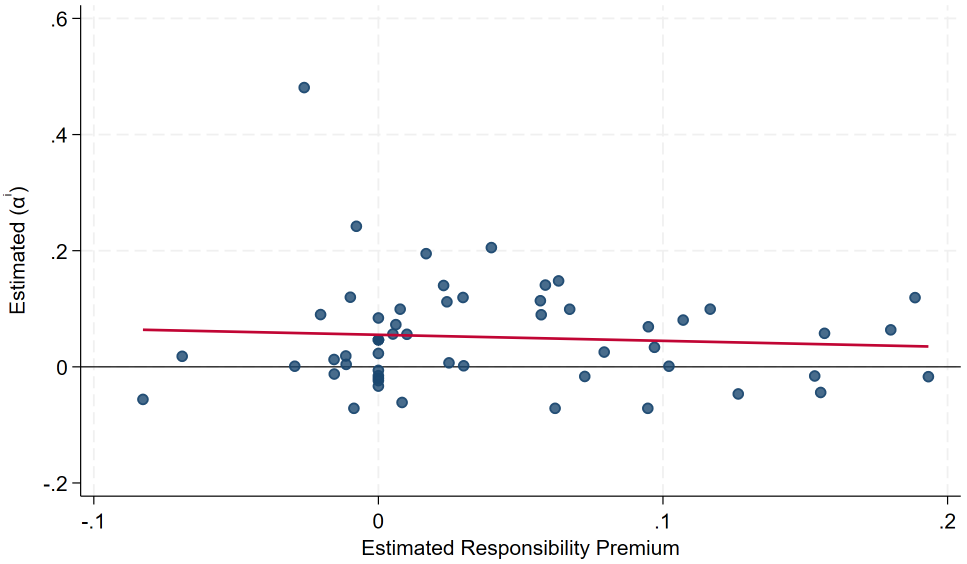


Figure 16: Correlation between Individual-Level Estimates of Amplification Parameter α^i and Responsibility Premium $\eta^{\mathrm{i}}_{\mathrm{resp}} - \eta^{\mathrm{i}}_{\mathrm{back}}$. α^i is estimated for a sample of 68 subjects who produced sufficient behavioural variation in `BothLow` and `BothHigh`. The red line corresponds to a linear fit from a regression of α^i on $\eta^{\mathrm{i}}_{\mathrm{resp}} - \eta^{\mathrm{i}}_{\mathrm{back}}$.

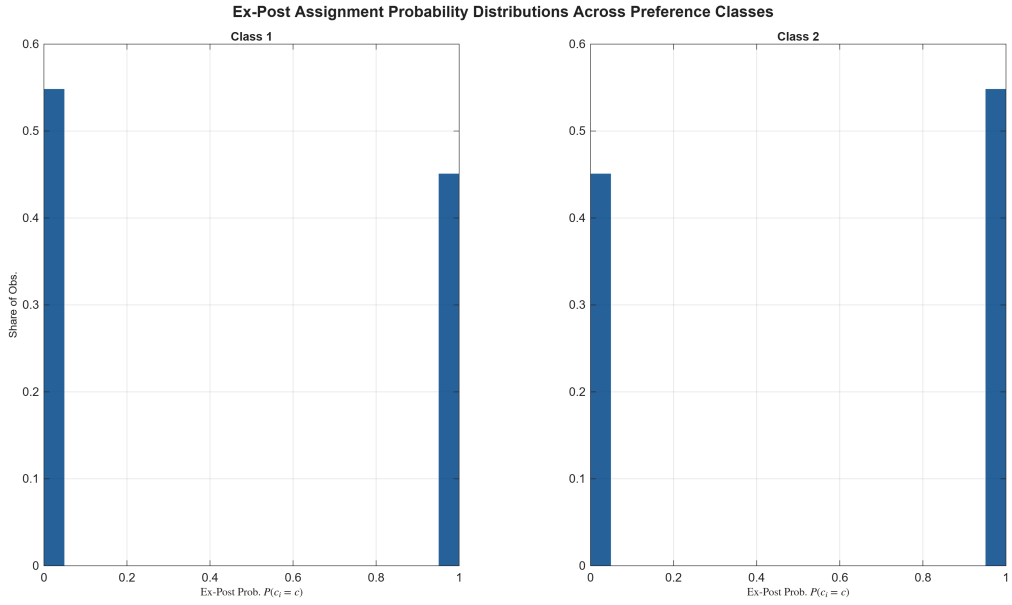


Figure 17: Finite Mixture Estimations: Ex-Post Class-Assignment Probabilities of Subjects in the Estimated Two-Type Model