Bartling, Björn; Cappelen, Alexander W.; Hermes, Henning; Skivenes, Marit; Tungodden, Bertil

Working Paper

Free to fail? Paternalistic preferences in the United States

Working Paper, No. 436

Provided in Cooperation with:
Department of Economics, University of Zurich

Suggested Citation: Bartling, Björn; Cappelen, Alexander W.; Hermes, Henning; Skivenes, Marit; Tungodden, Bertil (2025) : Free to fail? Paternalistic preferences in the United States, Working Paper, No. 436, University of Zurich, Department of Economics, Zurich, https://doi.org/10.5167/uzh-233707

This Version is available at:
https://hdl.handle.net/10419/322293

University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 436

Free to Fail? Paternalistic Preferences in the United States

Björn Bartling, Alexander W. Cappelen, Henning Hermes, Marit Skivenes and Bertil Tungodden

Revised version, May 2025

Free to Fail? Paternalistic Preferences in the United States*

Björn Bartling, Alexander W. Cappelen, Henning Hermes, Marit Skivenes & Bertil Tungodden

May 20, 2025

Abstract

This paper examines paternalistic preferences in large-scale experiments in the U.S. Participants decide whether to intervene to prevent a stakeholder, mistaken about their options, to make a choice that is misaligned with their preferences. We find that the willingness to intervene strongly depends on the nature of the paternalistic intervention: only a minority implements a hard intervention that limits the freedom to choose, while a majority implements a soft intervention that provides information without restricting the choice set. Based on a theoretical framework, we estimate that about half of the participants are welfarists, while a third are libertarian paternalists.

JEL classifications: C91, C93, D69, D91
Keywords: paternalism, libertarian paternalism, welfarism, freedom to choose

*We gratefully acknowledge funding from the Norwegian Research Council through its Centers of Excellence Scheme project 262675 (FAIR) and Research Grants 262636 and 325134, ERC Consolidator Grant 724460 and ERC Advanced Grant 788433. This study falls under the FAIR-The Choice Lab umbrella IRB approval for non-invasive survey experiments (NHH-IRB 13/20 and NHH-IRB 31/21). It was preregistered in the AEA RCT Registry (AEARCTR-0004630) and administered by FAIR-The Choice Lab. Camilla Allocchio provided excellent research support.

1 Introduction

People sometimes make choices that harm their welfare, creating opportunities for paternalistic interventions (Camerer, Issacharoff, Loewenstein, O'Donoghue, and Rabin, 2003; Thaler and Sunstein, 2003). The extent to which such opportunities should be utilized is a major debate concerning the relationship between the state and its citizens (Dworkin, 1972; Arneson, 1980; Le Grand and New, 2015). Two normative questions arise: First, is it acceptable to restrict an individual's freedom to choose in order to promote their welfare? Second, who should judge whether an intervention enhances the welfare of an individual? These questions have shaped an extensive normative literature, both in economics and the social sciences more broadly (Nussbaum, 2001; Sen, 2002; Kaplow and Shavell, 2006; Sunstein and Thaler, 2008; Hausman and McPherson, 2009; Le Grand and New, 2015).

To explore people's paternalistic preferences, we implemented a novel experimental design in two large-scale studies with 14,000 U.S. participants in the role of a third-party spectator, who can make an intervention decision that is consequential for another individual, the stakeholder. Study 1 examines people's views on the first normative question by investigating whether a concern for the stakeholder's freedom to choose affects the spectators' willingness to intervene. Study 2 examines people's views on the second normative question by eliciting spectators' evaluation of the stakeholder's welfare. The experimental design captures the key characteristics of situations that create opportunities for paternalistic interventions: a stakeholder is about to make a choice that is not aligned with their preferences, and a spectator can intervene to ensure that they receive their preferred option.

In Study 1, each spectator is matched with a stakeholder eligible for a monetary bonus. There are two bonus options in the stakeholder's choice set, a safe option and a risky option. Absent an intervention, the stakeholder will make their choice between the two options in a non-transparent choice environment. Our experiment creates a situation in which stakeholders, who prefer the safe option, mistakenly favor the risky option because of the non-transparent choice environment. In one treatment arm, spectators can implement a *hard* intervention, removing the stakeholder's freedom to choose by assigning the safe option. In another treatment arm, spectators can implement a *soft* intervention, providing information without restricting the freedom to choose. Both interventions differ only in whether they restrict the freedom to choose, such that the experimental design allows us to causally identify whether the nature of intervention—hard or soft—is an important factor in people's paternalistic preferences.

Study 1 includes a second treatment dimension that varies the reason why the stakeholder perceives, within the non-transparent choice environment, the risky option to be more attractive than it actually is. The literature on social preferences documented that the willingness to redistribute depends on the extent to which individuals are seen

as *responsible* for their situation (Konow, 2000; Fong, 2001; Cappelen, Drange Hole, Sørensen, and Tungodden, 2007; Alesina, Stantcheva, and Teso, 2018; Almås, Cappelen, and Tungodden, 2020). Likewise, the willingness to make a paternalistic intervention may depend on whether individuals are seen as responsible for making choices that are not aligned with their own preferences. We therefore study the role of the source of the stakeholder's mistake for the spectator's willingness to intervene. In one treatment arm, the spectator is informed that the stakeholder is mistaken about the choice set because they made an incorrect calculation, which we refer to as a situation with *internal* responsibility. In another treatment arm, the spectator is informed that the stakeholder is mistaken about the choice set because they received incorrect information, which we refer to as a situation with *external* responsibility.

Study 1 provides three main findings. First, we document that, in line with libertarian paternalism, the nature of a paternalistic intervention is a major causal determinant of the spectators' willingness to compete. Only about a third of the spectators are willing to implement the hard intervention that removes the stakeholder's freedom to choose, while a large majority of about 85 percent of the spectators are willing to implement the soft intervention that does not restrict the stakeholder's choice set. Second, we find that the source of the stakeholder's mistake is of minor importance for the spectators' willingness to intervene. Third, the heterogeneity analysis shows that the estimated treatment effects are robust across different subgroups of the general population in the United States.

We develop a theoretical framework to interpret our findings, formalizing two paternalistic positions: libertarian paternalism and welfarism. Libertarian paternalism justifies interventions that preserve freedom to choose and promote welfare, as judged by individuals themselves. This approach advocates interventions that manipulate the choice architecture without restricting choice, like defaults or information provision, nudging people towards choices aligned with their own preferences (Sunstein and Thaler, 2008). It has received considerable attention by policy makers and business leaders, as evidenced by the many behavioral insights teams established by governments and corporations across the world (OECD, 2017; DellaVigna and Linos, 2022; List, Rodemeier, Roy, and Sun, 2023). Welfarism, by contrast, assesses policies purely on their welfare impact (Sen, 2002). It accepts limiting choice sets if it improves welfare and generally aligns with libertarian paternalism's focus on individual preferences as a welfare basis (Kaplow and Shavell, 2000, 2001). However, welfarism can also incorporate alternative welfare conceptions that do not rely on satisfying an individual's own preferences (Nussbaum, 2001; Sen, 2002; Hausman and McPherson, 2009; Le Grand and New, 2015).

In Study 2, we estimate the prevalence of paternalistic types in the population by manipulating the nature of the intervention (as in Study 1) and by determining whether a spectator's welfare evaluation aligns with the stakeholder's preferences. Study 2 closely replicates the empirical pattern found in Study 1, confirming across two large

U.S. samples that the nature of an intervention significantly influences spectators' willingness to intervene. Further, we find that 70 percent of spectators make welfare evaluations aligned with the stakeholder's preferences. Based on these findings and our theoretical framework, we estimate that about 50 percent of spectators are welfarists, while about a third are libertarian paternalists. Together, these two paternalistic types can account for the behavior of most spectators across both studies.

These findings shed light on why soft interventions have received considerable attention in policy debates (OECD, 2017). Notably, the observation that the vast majority of the U.S. population is willing to implement soft interventions should not be interpreted as evidence of most U.S. citizens being libertarian paternalists. A large part of the support for the soft intervention in our studies comes from welfarists who are willing to implement both the soft and the hard intervention. Hence, the popularity of soft interventions can be explained by these interventions attracting support both from welfarists who respect the preferences of the stakeholder and from libertarian paternalists. In the same way, resistance to hard interventions that restrict people's choice sets may not only be driven by libertarian paternalists who respect people's freedom to choose, but may, as in our study, also reflect that a significant share of people are welfarists who believe that the hard intervention does not promote the welfare of the stakeholder. An interesting implication of the estimated prevalence of the different paternalistic types is that the libertarian paternalists are part of the majority coalition on the acceptability of the soft intervention and also part of the majority coalition on the non-acceptability of the hard intervention in our study. As a result, even though we estimate libertarian paternalists to comprise only about a third of the U.S. population, they may trigger both political support for implementing soft interventions and political resistance against hard interventions.

The paper contributes to the growing literature on paternalism by being the first experimental study to examine the role of the nature of a paternalistic intervention in shaping people's willingness to intervene. Several survey-based studies have shown that a majority of the population in various countries approve of a broad range of soft interventions (Diepeveen, Ling, Suhrcke, Roland, and Marteau, 2013; Reisch and Sunstein, 2016; Evers, Marchiori, Junghans, Cremers, and De Ridder, 2018; Sunstein, Reisch, and Rauber, 2018), even though there is some resistance to soft interventions that are considered to be manipulative (Jung and Mellers, 2016; Tannenbaum, Fox, and Rogers, 2017; Arad and Rubinstein, 2018). However, these studies do not examine the extent to which support for such policies is driven by people valuing others' freedom to choose or by a general preference for paternalistic interventions. We address this gap by providing evidence based on consequential choices made by large-scale U.S. samples.

Another strand of the literature considers how people conceptualize the welfare of others when considering paternalistic interventions. Ambuehl, Bernheim, and Ockenfels (2021) show that spectators project their own time preferences onto stakeholders and

are willing to restrict the stakeholders' choice sets by removing impatient options. Relatedly, in a field experiment, Kiessling, Chowdhury, Schildberg-Hörisch, and Sutter (2021) find that parents interfere in their children's intertemporal decision-making to mitigate their children's impatience. In the context of charitable donations, Jacobsson, Johannesson, and Borgquist (2007) and Gangadharan, Grossman, Jones, and Leister (2018) argue that altruism can be motivated paternalistically, by documenting a preference for in-kind donations over cash. The experimental design in the present study contributes to this line of research by showing the extent to which people respect the preferences of others when they have complete information about those preferences.

More broadly, our paper contributes to the literature on people's social preferences and the heterogeneity of these preferences among individuals (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Konow, 2000; Andreoni and Miller, 2002; Cappelen et al., 2007; Bellemare, Kröger, and van Soest, 2008; Dufwenberg, Heidhues, Kirchsteiger, Riedel, and Sobel, 2011; Erkal, Gangadharan, and Nikiforakis, 2011; Durante, Putterman, and Weele, 2014; Almås et al., 2020). We show substantial heterogeneity in people's paternalistic preferences and formalize two paternalistic types that capture this heterogeneity. Further, in contrast to the literature showing that the idea of personal responsibility is of great importance for inequality acceptance (Konow, 2000; Fong, 2001; Cappelen et al., 2007), we find that the willingness to intervene does not depend on whether individuals are perceived as responsible for making choices that are not aligned with their own preferences. Finally, the paper contributes to the literature on the intrinsic value of decision rights, power, and self-determination (Deci and Ryan, 1985; Fehr, Herz, and Wilkening, 2013; Bartling, Fehr, and Herz, 2014; Pikulina and Tergiman, 2020; Freundt, Herz, and Kopp, 2023), by providing evidence that people value not only their own autonomy but also respect others' freedom to choose.

The paper is organized as follows. Section 2 presents Study 1, followed by Section 3, introducing the theoretical framework. Section 4 discusses Study 2. The paper concludes in Section 5. The Online Appendix includes alternative regression models, adjustments for multiple hypothesis testing, further details on experimental procedures and instructions, and outlines the minor deviations from the pre-analysis plans.

2 Study 1

2.1 Experimental Design

The experiment has two types of participants: spectators and stakeholders. The spectators make intervention decisions that are consequential for the stakeholders. Our interest is in the spectators' intervention decisions, and the sole function of the stakeholders is to render the spectators' decisions consequential.[1]

[1] Online Appendix B presents the experimental instructions for both spectators and stakeholders.

Context. Each spectator is matched with a stakeholder who will receive a bonus payment. There are two bonus options, a safe payment of USD 4 and a risky payment of USD 10 or USD 0 with equal probability. Absent an intervention, the stakeholder will make a choice between the two bonus options in a non-transparent choice environment. The non-transparent choice environment leads the stakeholder to be mistaken about the choice set. Specifically, the stakeholder is mistaken about the odds of the risky option and, as a consequence, the stakeholder prefers the risky option over the safe option. However, the stakeholder would prefer the safe option over the risky option if they were not mistaken about the odds of the risky option.

More formally, let s denote the option with the safe payment of USD 4, and r denote the option with the risky payment of USD 10 or USD 0 with equal probability. In the non-transparent choice environment, the stakeholder mistakenly believes that the choice is not between s and r, but between s and a different risky option, which we refer to as $\tilde{r} \neq r$. The stakeholder's preference ranking is given by $\tilde{r} \succ s \succ r$.[2]

The spectator is fully informed about the stakeholder's preference ranking, and the experimental design relies on the minimal assumption that the spectator believes the stakeholder acts in accordance with their preferences. A stakeholder will choose the risky option in the non-transparent choice environment because, in this case, the stakeholder mistakenly believes that the choice is between s and $\tilde{r}$. Consequently, absent an intervention, the stakeholder ends up with the non-preferred option r. The spectator is given the opportunity to intervene to ensure that the stakeholder ends up with their preferred option s.

This context captures the key characteristics of situations that create opportunities for paternalistic interventions: a stakeholder is about to make a choice misaligned with their preferences, and a spectator can intervene to help them secure their preferred option.

Nature of Intervention. The spectators are randomly assigned either to treatments where they can implement a *hard* intervention or to treatments where they can implement a *soft* intervention. The hard intervention removes the stakeholder's freedom to choose, and the stakeholder is allocated the safe option s. In contrast, the soft intervention retains the stakeholder's freedom to choose but provides them with correct information about the odds of the risky option. As a result, the stakeholder will know that the choice is between s and r and, consequently, chooses s.

It follows that the outcome of intervening and the outcome of not intervening does not depend on the nature of the intervention: if the spectator intervenes, hard or soft, the stakeholder ends up with their preferred safe option s; if the spectator does not intervene, hard or soft, the stakeholder ends up with their non-preferred risky option r.

[2]Online Appendix B.1 details how the experimental design allows us to identify stakeholders with this preference ranking.

Hence, if the spectators' willingness to intervene only depends on the outcome of an intervention, the share of spectators that implement the hard intervention would be equal to the share of spectators that implement the soft intervention. The experimental design thus allows us to identify whether the nature of the intervention matters for the spectators' willingness to intervene.

Source of Mistake. We also study the role of the source of the stakeholder's mistake. Spectators are randomly assigned either to treatments where the source of mistake is *internal* or to treatments where the source of mistake is *external*. In treatments where the source of mistake is internal, the spectator is informed that they are matched to a stakeholder who had to calculate the odds of the risky option and made a mistake in the calculations. In treatments where the source of mistake is external, the spectator is informed that they are matched to a stakeholder who was unlucky and received an incorrect signal about the odds of the risky option.[3] Since the source of mistake does not affect the choice of the stakeholder in the non-transparent environment (for both sources of mistake, $\tilde{r} \succ s \succ r$), it does not affect the outcomes of intervening or not intervening. Consequently, if spectators' willingness to intervene does not depend on the source of mistake, the share of spectators intervening should be equal for internal and external mistakes. The experimental design thus allows us to identify whether the source of a stakeholder's mistake matters for spectators' willingness to intervene.

Treatment Design. We implemented a factorial 2×2 between-subjects design: *Hard*×*Internal*, *Hard*×*External*, *Soft*×*Internal*, and *Soft*×*External*. Figure 1 provides and overview of the stages of the experiment and when the treatment manipulations come into play.

[Figure 1 about here]

2.2 Participants

Spectators. The spectators were recruited from the general U.S. population through the survey provider Dynata. We sampled 8,004 spectators in August 2019, based on quotas for gender, age, education, income, and region, to match a representative sample of the population (aged 18 or older). Spectators had to pass an attention filter before being randomized with equal probability to one of the four treatments. Each spectator made a single intervention decision. The spectators were informed that one out of five spectator decisions would be randomly selected and implemented.

[3] All participants were only provided with truthful information, see Online Appendix B.1. Specifically, the spectators were informed that the stakeholders knew the average of all signals sent was correct, even if the signal they received might be incorrect.

We elicited demographic background characteristics, including gender, age, education, and income. Further, since the intervention decision is made in the domain of risk, we measured the spectators' willingness to take risks by eliciting their self-assessment on an 11-point scale ranging from "Completely unwilling to take risks" to "Very willing to take risks". Finally, spectators could self-identify as "Republican," "Democrat," or "Independent/Third Party," and provide their agreement with the following statements: "People sometimes make choices that harm their own well-being" and "The government can sometimes improve its citizens' well-being by restricting their freedom of choice" on a seven-point scale ranging from "fully disagree" to "fully agree."

Our sample closely mirrors the population statistics with respect to gender and age, but contains a slightly lower share with low education and with a household income of at least USD 150,000.[4] Regarding political affiliation, 29% self-identify as Republicans, 33% as Democrats, and 28% as Independents/Third Party, while 10% did not report a political affiliation.

Stakeholders. The stakeholders were recruited on AMTurk. Only stakeholders who prefer the safe option in the transparent choice environment and the risky option in the non-transparent choice environment were matched to spectators.

2.3 Empirical Strategy

To examine how the nature of the intervention and the source of mistake causally affect the spectators' willingness to intervene, we use the following empirical specification:[5]

$$I_i = \beta_0 + \beta_1 S_i + \beta_2 E_i + \beta_3 S_i E_i + \gamma X_i + \varepsilon_i \tag{1}$$

The dependent variable I_i is an indicator for whether spectator i intervenes. Treatment $Hard \times Internal$ is the omitted category. S_i is an indicator for spectator i being in a treatment with a soft intervention, E_i is an indicator for spectator i being in a treatment with the external source of mistake, $S_i E_i$ is the interaction between S_i and E_i, X_i is a vector of background characteristics, and ε_i is an idiosyncratic error term. X_i includes political orientation, willingness to take risks, education, income, age, and gender. In the analysis, the background characteristics are defined by the following indicator variables: Republican indicates whether a spectator identifies as Republican or non-Republican. High Risk Taking, High Education, High Income, and High Age indicate whether a spectator is above or below the median of the respective characteristic in the sample. Female indicates whether a spectator is female or male. We estimate the models with and without the vector of background characteristics.

[4] Online Appendix Table A1 shows the demographic characteristics of the sample and compares it to the U.S. population. Table A2 shows that the sample is balanced across treatments.

[5] In the appendix, we show that all our results prevail in Probit regressions.

The coefficient β_1 provides an estimate of the causal effect of the nature of the intervention on the spectators' willingness to intervene. The coefficient β_2 provides an estimate of the causal effect of the source of mistake on the spectators' willingness to intervene. The coefficient β_3 provides an estimate of the interaction effect between the nature of the intervention and the source of mistake on the spectators' willingness to intervene. We also estimate the causal effect of the nature of the intervention when pooling the treatments with the hard intervention (*Hard*×*Internal* and *Hard*×*External*) and the treatments with the soft intervention (*Soft*×*Internal* and *Soft*×*External*), and we estimate the causal effect of the source of mistake when pooling the treatments with the internal source of mistake (*Hard*×*Internal* and *Soft*×*Internal*) and the treatments with the external source of mistake (*Hard*×*External* and *Soft*×*External*).

In the heterogeneity analysis, we study the average causal effect of the nature of the intervention and the source of mistake in different subgroups when pooling the respective treatments. In this analysis, we use the following specification:

$$I_i = \beta_0 + \beta_1 T_i + \beta_2 x_i + \beta_3 T_i x_i + \varepsilon_i \tag{2}$$

where x_i indicates a single background characteristic of spectator i, and $T_i x_i$ is the interaction between x_i and the treatment indicator $T_i = S_i, E_i$. Equation (2) is estimated separately for x_i indicating political orientation, willingness to take risks, education, income, gender, and age (single interaction model). In the analysis, we also estimate a model that jointly includes x_i and the interaction term $T_i x_i$ for each background characteristic (joint interaction model).

2.4 Results

Figure 2 shows the share of spectators that intervene in treatments *Hard*×*Internal* and *Soft*×*Internal* on the left, and the share of spectators that intervene in treatments *Hard*×*External* and *Soft*×*External* on the right.

[Figure 2 about here]

We observe that only about a third of the spectators implement the hard intervention, both when the source of mistake is internal (33.5 percent) and when the source of mistake is external (30.0 percent). Hence, independent of the source of mistake, the large majority of spectators decide not to restrict the stakeholder's freedom to choose, even though the hard intervention would ensure that the stakeholder ends up with their preferred safe option. In contrast, a large majority of the spectators implement the soft intervention, which preserves the stakeholder's freedom to choose while ensuring

that the stakeholder ends up with their preferred safe option (internal: 85.9 percent, external: 87.5 percent).

Table 1 reports regression analysis.[6] In Column (1), we estimate the average causal effect of the nature of the intervention on the spectators' willingness to intervene: the share of spectators that implement the soft intervention is 55.0 percentage points higher than the share of spectators that implement the hard intervention ($p < 0.01$). Column (2) shows that the estimated effect is robust to the inclusion of the spectators' background characteristics. We further note that the estimated coefficients for the background characteristics are small and in most cases not significant.

[Table 1 about here]

Columns (3) and (4) estimate the model with an interaction variable between the nature of the intervention and the source of mistake. We find only a small interaction effect: the estimated difference between the share of spectators that implement the hard intervention and the share of spectators that implement the soft intervention is 5.1 percentage points larger when the source of mistake is external rather than internal ($p < 0.01$). Consequently, the estimated treatment effect of manipulating the nature of the intervention is large both when the source of mistake is internal (52.4 percent, $p < 0.01$) and when it is external (57.5 percent, $p < 0.01$).

Result 1: *The nature of the intervention has a substantial causal effect on the spectators' willingness to intervene, both when the source of the stakeholder's mistake is internal and when it is external.*

We now turn to an analysis of the causal effect of the source of the stakeholder's mistake on the spectators' willingness to intervene. Columns (3) and (4) show that the source of mistake does not have a large effect on the willingness to implement the hard intervention or the soft intervention, and Columns (5) and (6) show that, across the two interventions, there is no significant estimated effect of the source of mistake on the spectators' willingness to intervene.

Result 2: *The source of the stakeholder's mistake does not have a substantial causal effect on the spectators' willingness to intervene, irrespective of the nature of the intervention.*

Our large-scale general population sample allows us to study whether different subgroups of the population—defined by political orientation, willingness to take risks,

[6]The analysis follows the pre-analysis plans, registered separately for Study 1 and Study 2, with only very minor deviations discussed in Online Appendix C.

education, income, age, and gender—differ in their intervention decisions. In Online Appendix Tables A6 and A7, we show that the estimated causal effect of the nature of the intervention on the spectators' willingness to intervene is not significantly different across subgroups when correcting for multiple hypothesis testing, with the exception of education. The estimated treatment effect of the nature of the intervention is 6.0 percentage points larger for spectators with high education than for spectators with low education ($p < 0.01$). In Online Appendix Tables A8 and A9, we show that the estimated causal effect of the source of mistake on the spectators' willingness to intervene is not significantly different across subgroups.

Further, the analysis in Tables A6 to A9 shows that the estimated effect of the nature of the intervention is large and highly significant in all subgroups ($p < 0.01$), while the estimated effect of the source of mistake is small and insignificant in all subgroups.

Result 3: *There are generally only small differences in the intervention decisions across subgroups. In all subgroups, the nature of the intervention has a substantial average causal effect on the spectators' willingness to intervene, while the source of the stakeholder's mistake does not.*

3 Theoretical Framework

In this section, we provide a theoretical framework to guide the interpretation of our results. Let θ^H denote the share of spectators that implement the hard intervention and θ^S the share of spectators that implement the soft intervention. Study 1 establishes, both for the internal and the external source of mistake, the following empirical pattern:

$$0 < \hat{\theta}^H < \hat{\theta}^S < 1, \tag{3}$$

where $\hat{\theta}^H$ and $\hat{\theta}^S$ denote the estimated shares. This raises three questions about the spectators' intervention decisions:

- Why do some spectators implement the hard intervention?
- Why do more spectators implement the soft than the hard intervention?
- Why do some spectators not implement the soft intervention?

3.1 Spectators' Preferences

We assume that a spectator's preferences are defined over the stakeholder's welfare and the stakeholder's freedom to choose.

Let $W(b)$ denote a spectator's evaluation of the stakeholder's welfare, which is determined by the bonus option, $b \in \{s, r\}$, that the stakeholder ends up with. We assume that either $W(s) > W(r)$ or $W(s) < W(r)$. Let $U(b)$ represent the preference ranking of the stakeholder, with $U(s) > U(r)$ for all stakeholders in the experiment. A spectator's welfare evaluation is aligned with the stakeholder's preferences if and only if $W(b) = U(b)$, which in the experiment would entail that $W(s) > W(r)$. Let θ^A denote the share of spectators whose welfare evaluations are aligned with the stakeholder's preferences.

Let $F(c)$ denote a spectator's evaluation of the stakeholder's freedom to choose, which is determined by the stakeholder's choice environment, $c \in \{c^{+t}, c^{+nt}, c^{-}\}$, where c^{+t} denotes a transparent choice environment, c^{+nt} a non-transparent choice environment, and c^{-} an environment in which the stakeholder has no choice. We assume that $F(c^{+t}) \geq F(c^{+nt}) > F(c^{-})$.

3.2 Spectators' Intervention Decisions

A spectator can choose to intervene, i, or not to intervene, ni. In the experiment, $b(i) = s$ and $b(ni) = r$, both in treatment *Hard* and in treatment *Soft*. Hence, the welfare consequences of intervening and of not intervening are the same in the two treatments. Further, $c = c^{-}$ if a spectator intervenes in treatment *Hard* and $c = c^{+t}$ if a spectator intervenes in treatment *Soft*. In both treatments, $c(ni) = c^{+nt}$. The stakeholder's freedom to choose is strictly reduced by intervening in treatment *Hard* and is weakly increased by intervening in treatment *Soft*.

We make the following two minimal assumptions about a spectator's intervention decisions:

A1. A spectator intervenes if $W(b(i)) > W(b(ni))$ and $F(c(i)) \geq F(c(ni))$.

A2. A spectator does not intervene if $W(b(i)) < W(b(ni))$ and $F(c(i)) \leq F(c(ni))$.

The two assumptions imply that a spectator does not make dominated intervention decisions: a spectator who considers that an intervention strictly increases the stakeholder's welfare and at least weakly increases the stakeholder's freedom to choose will intervene, and a spectator who considers that an intervention strictly decreases the stakeholder's welfare and at least weakly decreases the stakeholder's freedom to choose will not intervene.

We can now make the following observation:

Observation 1: Assumptions A1 and A2 imply that $\theta^H \leq \theta^A \leq \theta^S$.

Proof. (i) Consider a spectator who implements the hard intervention. By A2, $W(s) > W(r)$. It follows that $\theta^H \leq \theta^A$. (ii) Consider a spectator who does not implement the soft intervention. By A1, $W(s) < W(r)$. It follows that $\theta^A \leq \theta^S$. □

The empirical pattern observed in Study 1, $0 < \hat{\theta}^H < \hat{\theta}^S < 1$, is consistent with Observation 1. However, Study 1 does not provide us with a measure of θ^A, the share of spectators whose welfare evaluations are aligned with the stakeholder's preferences. To provide a stricter test of the theoretical framework, we implement Study 2, which comprises a treatment that provides us with an estimate of θ^A, along with estimates of θ^H and θ^S. Study 2 also allows us to study the prevalence of the main paternalistic types in the normative literature, which we now turn to.

3.3 Paternalistic Types

Within the theoretical framework, libertarian paternalism and welfarism can be formalized as follows:

Libertarian Paternalist. A libertarian paternalist has aligned preferences, $W(b) = U(b)$, and intervenes if and only if $F(c(i)) \geq F(c(ni))$ and $W(b(i)) > W(b(ni))$.

Welfarist. A welfarist intervenes if and only if $W(b(i)) > W(b(ni))$.

The two paternalistic types satisfy assumptions A1 and A2.

A libertarian paternalist intervenes if and only if the intervention preserves the stakeholder's freedom to choose and strictly increases the stakeholder's welfare, with the welfare evaluation being aligned with the stakeholder's preferences. Hence, a libertarian paternalist does not implement any intervention that reduces the stakeholder's freedom to choose, irrespective of how the intervention affects the stakeholder's welfare. A welfarist intervenes if and only if the intervention strictly increases the stakeholder's welfare, irrespective of how the intervention affects the stakeholder's freedom to choose. The welfare evaluation of a welfarist may or may not be aligned with the stakeholder's preferences.

It follows that:

- Welfarists whose welfare evaluations are aligned with the stakeholder's preferences implement both the hard and the soft intervention.
- Libertarian paternalists implement the soft but not the hard intervention.

- Welfarists whose welfare evaluations are not aligned with the stakeholder's preferences neither implement the hard nor the soft intervention.

We now consider how the theoretical framework can be used to study the prevalence of the two paternalistic types.

3.4 Prevalence of Paternalistic Types

Let σ^{LP} denote the share of spectators that are libertarian paternalists, σ^{W_a} the share of spectators that are welfarists whose welfare evaluations are aligned with the stakeholder's preferences, and $\sigma^{W_{na}}$ the share of spectators that are welfarists whose welfare evaluations are not aligned with the stakeholder's preferences.

We establish the following observation:

Observation 2: Assumptions A1 and A2 imply that (i) the share of welfarists whose welfare evaluations are aligned with the stakeholder's preferences is given by $\sigma^{W_a} = \theta^H$, (ii) the share of welfarists whose welfare evaluations are not aligned with the stakeholder's preferences is given by $\sigma^{W_{na}} = 1 - \theta^S$, (iii) the share of libertarian paternalists is given by $\sigma^{LP} = \theta^A - \theta^H$, and (iv) $\sigma^{W_a} + \sigma^{W_{na}} + \sigma^{LP} \leq 1$.

Proof. (i) Consider a spectator who implements the hard intervention. By A2, $W(s) > W(r)$. By A1, the spectator implements the soft intervention. Hence, the spectator is a welfarist whose welfare evaluation is aligned with the stakeholder's preferences. It follows that $\sigma^{W_a} = \theta^H$. (ii) Consider a spectator who does not implement the soft intervention. By A1, $W(r) > W(s)$. By A2, the spectator does not implement the hard intervention. Hence, the spectator is a welfarist whose welfare evaluation is not aligned with the stakeholder's preferences. It follows that $\sigma^{W_{na}} = 1 - \theta^S$. (iii) Consider a spectator for whom it holds that $W(s) > W(r)$. By A1, the spectator implements the soft intervention. Hence, the spectator is either a welfarist (if the spectator implements the hard intervention) or a libertarian paternalist (if the spectator does not implement the hard intervention). It follows that $\sigma^{W_a} + \sigma^{LP} = \theta^A$. Taking into account (i), it follows that $\sigma^{LP} = \theta^A - \theta^H$. (iv) It follows from (i), (ii), and (iii) that $\sigma^{W_a} + \sigma^{W_{na}} + \sigma^{LP} = \theta^H + (1 - \theta^S) + (\theta^A - \theta^H) = 1 - \theta^S + \theta^A$. By A1, $\theta^A \leq \theta^S$. It follows that $\sigma^{W_a} + \sigma^{W_{na}} + \sigma^{LP} \leq 1$. □

It follows from Observation 2 that if *all* spectators are either welfarists or libertarian paternalists, then $\theta^A = \theta^S$. However, A1 and A2 allow for behavior that cannot be rationalized by either welfarists or libertarian paternalists: spectators who implement the soft intervention, even though their welfare evaluations are not aligned with the stakeholder's preferences. These spectators must (i) evaluate the stakeholder's freedom to choose to be strictly greater in a transparent choice environment than in a

non-transparent choice environment (otherwise, they would violate A2 by implementing the soft intervention) and (ii) consider the increase in the stakeholder's freedom to choose from implementing the soft intervention to outweigh what they evaluate to be a loss in the stakeholder's welfare. Taken together, these spectators cannot be welfarists because they care about the stakeholder's freedom to choose and they cannot be libertarian paternalists because their welfare evaluations are not aligned with the stakeholder's preferences. The share of such spectators is given by $\theta^S - \theta^A$. In principle, all spectators could agree with (i) and (ii), which would be the case if $0 = \theta^H = \theta^A < \theta^S = 1$.

We now turn to Study 2, which tests the theoretical framework (Observation 1) and estimates the shares of the paternalistic preference types (Observation 2) based on both the spectators' intervention decisions and their welfare evaluations.

4 Study 2

4.1 Experimental Design and Participants

Design. Study 2 replicates treatments *Hard*×*Internal* and *Soft*× *Internal* from Study 1 (referred to as treatments *Hard* and *Soft* in the following) and adds a treatment *Welfare* to directly elicit whether a spectator's welfare evaluations align with the stakeholders' preferences. The context of treatment *Welfare* is identical to that of treatments *Hard* and *Soft*, except that the spectator does not have the option to give the stakeholder the freedom to choose. Instead, the spectator must allocate either the preferred safe option or the non-preferred risky option to the stakeholder.

Participants. Spectators and stakeholders were recruited from the same populations as in Study 1, using the same procedures. Subjects who participated in Study 1 could not participate in Study 2. We sampled 6,033 spectators in January 2020.[7]

4.2 Empirical Strategy

To analyze the spectators' decisions in Study 2, we use the following empirical specification:

$$D_i = \beta_0 + \beta_1 H_i + \beta_2 S_i + \gamma X_i + \varepsilon_i \tag{4}$$

[7]Online Appendix Table A1 reports the sample characteristics, which closely resemble those of Study 1. Table A3 shows that the sample is largely balanced across treatments, with slightly fewer Republicans and slightly more spectators with higher education and income in treatment *Welfare*.

The dependent variable D_i is an indicator for whether spectator i intervenes in treatments *Hard* or *Soft*, or allocates the safe option in treatment *Welfare*. Treatment *Welfare* is the omitted category in the regression model. H_i is an indicator for spectator i being in treatment *Hard*, S_i is an indicator for spectator i being in treatment *Soft*, X_i is a vector of background characteristics (political orientation, willingness to take risks, education, income, age, and gender), and ε_i is an idiosyncratic error term. We estimate the model with and without the vector of background characteristics.

The regression model can be used to test Observation 1 ($\theta^H \leq \theta^A \leq \theta^S$). It follows from the regression model that $\hat{\theta}^H = \hat{\beta}_0 + \hat{\beta}_1$, $\hat{\theta}^S = \hat{\beta}_0 + \hat{\beta}_2$, and $\hat{\theta}^A = \hat{\beta}_0$. Hence, Observation 1 is rejected in the data if $\hat{\beta}_1 > 0$ or $\hat{\beta}_2 < 0$.

It follows from Observation 2 and the regression model that the estimated shares of libertarian paternalists and welfarists in our sample are given by: $\hat{\sigma}^{W_a} = \hat{\theta}^H = \hat{\beta}_0 + \hat{\beta}_1$, $\hat{\sigma}^{W_{na}} = 1 - \hat{\theta}^S = 1 - \hat{\beta}_0 - \hat{\beta}_2$, and $\hat{\sigma}^{LP} = \hat{\theta}^A - \hat{\theta}^H = -\hat{\beta}_1$. If all spectators are either libertarian paternalists or welfarists, then $\hat{\beta}_2 = 0$ and $\hat{\sigma}^{W_a} + \hat{\sigma}^{W_{na}} + \hat{\sigma}^{LP} = 1$.

4.3 Results

Figure 3 shows the share of spectators that intervene in treatments *Hard* and *Soft*, and the share that allocates the preferred safe option to the stakeholder in treatment *Welfare*. About a third of the spectators (35.4 percent) implement the hard intervention and the large majority of the spectators (82.5 percent) implement the soft intervention. Study 2 thus replicates Result 1 from Study 1: the nature of an intervention has a strong causal impact on the spectators' willingness to intervene. Moreover, Figure 3 shows that 69.8 percent of the spectators allocate the safe option to the stakeholder in treatment *Welfare*, while 30.2 percent of the spectators allocate the risky option to the stakeholder.

[Figure 3 about here]

Online Appendix Table A11 reports the corresponding regression analysis. The estimated share of spectators that intervene in treatment *Soft* is 12.7 percentage points higher than the estimated share of spectators that allocate the safe option to the stakeholder in treatment *Welfare* ($p < 0.01$). Further, we find that the estimated share of spectators that intervene in treatment *Hard* is 34.4 percentage points lower than the estimated share of spectators that allocate the safe option in treatment *Welfare* ($p < 0.01$). The treatment differences are virtually unaffected by the inclusion of the spectators' background characteristics.

These findings are aligned with Observation 1 ($\theta^H \leq \theta^A \leq \theta^S$). The estimated share of spectators that implement the hard intervention is strictly smaller than the estimated

share of spectators whose welfare evaluations are aligned with the stakeholder's preferences, and the estimated share of spectators that implement the soft intervention is strictly larger than the estimated share of spectators whose welfare evaluations are aligned with the stakeholder's preferences.

Result 4: *The spectators' intervention decisions are consistent with the theoretical framework, which provides evidence that the willingness to intervene is determined by how an intervention affects the stakeholder's welfare and freedom to choose.*

Based on the regression results, the estimated share of libertarian paternalists, $\hat{\sigma}^{LP}$, is 34.4 percent and the estimated share of welfarists, $\hat{\sigma}^{W_a} + \hat{\sigma}^{W_{na}}$, is 52.9 percent. The estimated share of welfarists whose welfare evaluations are aligned with the stakeholder's preferences, $\hat{\sigma}^{W_a}$, is 35.4 percent, and the estimated share of welfarists whose welfare evaluations are not aligned with the stakeholder's preferences, $\hat{\sigma}^{W_{na}}$, is 17.5 percent. Consequently, only the decisions of 12.7 percent of the spectators (given by $\hat{\theta}^S - \hat{\theta}^A$) cannot be rationalized the two paternalistic types. These spectators are not welfarists, as they value the stakeholder's freedom to choose, nor are they libertarian paternalists since their welfare evaluations are not aligned with the stakeholder's preferences.

Result 5: *The large majority of the spectators are either libertarian paternalists or welfarists: about a third of the spectators are estimated to be libertarian paternalists and about half of the spectators are estimated to be welfarists.*

Figure 4 shows that Observation 1 holds in all subgroups, see also Online Appendix Table A12. Moreover, we see that the spectator behavior is very similar across subgroups, with one exception: the share of spectators allocating the safe option in treatment *Welfare* is highest among below-median risk-takers and lowest among above-median risk-takers. This indicates that some spectators base welfare evaluations on their own risk preferences, which is consistent with the notion of "ideals-projective paternalism" in Ambuehl et al. (2021).

[Figure 4 about here]

5 Conclusions

The present study provides causal evidence on how the nature of an intervention—hard or soft—shapes people's willingness to intervene, in two independent large-scale samples of the U.S. population. Only about a third of the spectators implement a hard intervention that removes the stakeholder's freedom to choose, while a large majority

implement a soft intervention that provides information without restricting the choice set. We find that this result holds regardless of the stakeholder's responsibility for being mistaken about their choice set—whether the source of mistake is internal or external—and in all subgroups of the population.

We introduce a theoretical framework with two paternalistic types—libertarian paternalists and welfarists—and find that the behavior of the large majority of spectators can be rationalized by these types: about a third of the spectators are estimated to be libertarian paternalists and about half of the spectators are estimated to be welfarists. The estimated share of libertarian paternalists shows that a significant part of the U.S. population prefers leaving people "free to fail," rather than improving their welfare by restricting their freedom to choose. At the same time, we find great support for soft interventions that enable people to make choices in line with their preferences. The observed support for soft interventions in our study may represent an upper bound, as information provision may be regarded as less controversial than other soft interventions, such as default options and nudges, that have been criticized for being manipulative.

To assess whether the experimental context is relevant for the policy debate on paternalism in the U.S., we asked the spectators two general questions. First, we asked whether they agree that people sometimes make choices that are harmful to themselves. A large majority agreed, see upper panel of Online Appendix Figure A1. Second, we asked whether they agree that the government can sometimes improve people's lives by restricting their freedom to choose. We find sizable agreement with this view of the government but also substantial skepticism regarding the government's ability to improve people's lives with hard paternalistic policies, see lower panel of Figure A1. This skepticism could reflect general distrust in the government (Kuziemko, Norton, Saez, and Stantcheva, 2015). Republican spectators are more skeptical to government's ability to improve people's lives than Democratic spectators. In contrast, we do not find substantial political differences in spectator behavior in the experiment. Taken together, these findings suggest that political disagreements about paternalistic policies are more related to disagreements about the consequences of paternalistic interventions, which are controlled for in our experiment, than about fundamental differences in paternalistic preferences.

The experimental paradigm introduced in this paper can be used to explore further the nature of people's paternalistic preferences, including how they vary across domains and are influenced by stakes. Finally, an intriguing avenue for future research is to study cultural variation in paternalistic preferences and the extent to which this variation explains cross-country differences in attitudes toward paternalistic policies (Sunstein et al., 2018). Such policies represent a fundamental dimension of the relationship between the state and its citizens, making it essential to understand how they are justified and whether they align with people's paternalistic preferences.

References

Alesina, Alberto, Stefanie Stantcheva, and Edoardo Teso (2018). "Intergenerational mobility and preferences for redistribution," *American Economic Review*, 108(2): 521–554.

Almås, Ingvild, Alexander W. Cappelen, and Bertil Tungodden (2020). "Cutthroat capitalism versus cuddly socialism: Are Americans more meritocratic and efficiency-seeking than Scandinavians?" *Journal of Political Economy*, 128(5): 1753–1788.

Ambuehl, Sandro, Douglas B. Bernheim, and Axel Ockenfels (2021). "What motivates paternalism? an experimental study," *American Economic Review*, 111(3): 787–830.

Andreoni, James and John Miller (2002). "Giving according to GARP: An experimental test of the consistency of preferences for altruism," *Econometrica*, 70(2): 737–753.

Arad, Ayala and Ariel Rubinstein (2018). "The people's perspective on libertarian-paternalistic policies," *The Journal of Law and Economics*, 61(2): 311–333.

Arneson, Richard J (1980). "Mill versus paternalism," *Ethics*, 90(4): 470–489.

Bartling, Björn, Ernst Fehr, and Holger Herz (2014). "The intrinsic value of decision rights," *Econometrica*, 82(6): 2005–2039.

Bellemare, Charles, Sabine Kröger, and Arthur van Soest (2008). "Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities," *Econometrica*, 76(4): 815–839.

Bolton, Gary E. and Axel Ockenfels (2000). "ERC: A theory of equity, reciprocity, and competition," *American Economic Review*, 90(1): 166–193.

Camerer, Colin F., Samuel Issacharoff, George Loewenstein, Ted O'Donoghue, and Matthew Rabin (2003). "Regulation for conservatives: Behavioral economics and the case for asymmetric paternalism," *University of Pennsylvania law review*, 151(3): 1211–1254.

Cappelen, Alexander W., Astri Drange Hole, Erik Ø. Sørensen, and Bertil Tungodden (2007). "The pluralism of fairness ideals: An experimental approach," *American Economic Review*, 97(3): 818–827.

Deci, Edward L. and Richard M. Ryan (1985). *Intrinsic Motivation and Self-Determination in Human Behavior*, Springer.

DellaVigna, Stefano and Elizabeth Linos (2022). "Rcts to scale: Comprehensive evidence from two nudge units," *Econometrica*, 90: 81–116.

Diepeveen, Stephanie, Tom Ling, Marc Suhrcke, Martin Roland, and Theresa M. Marteau (2013). "Public acceptability of government intervention to change health-related behaviours: a systematic review and narrative synthesis," *BMC Public Health*, 13(756).

Dufwenberg, Martin, Paul Heidhues, Georg Kirchsteiger, Frank Riedel, and Joel Sobel (2011). "Other-regarding preferences in general equilibrium," *Review of Economics and Statistics*, 78(2): 613–639.

Durante, Ruben, Louis Putterman, and Joël Weele (2014). "Preferences for redistribution and perception of fairness: An experimental study," *Journal of the European Economic Association*, 12(4): 1059–1086.

Dworkin, Gerald (1972). "Paternalism," *the Monist*: 64–84.

Erkal, Nisvan, Lata Gangadharan, and Nikos Nikiforakis (2011). "Relative earnings and giving in a real-effort experiment," *American Economic Review*, 101(7): 3330–48.

Evers, C., D.R. Marchiori, A. F. Junghans, J. Cremers, and D. T. D. De Ridder (2018). "Citizen approval of nudging interventions promoting healthy eating: the role of intrusiveness and trustworthiness," *BMC Public Health*, 18.

Fehr, Ernst, Holger Herz, and Tom Wilkening (2013). "The lure of authority: Motivation and incentive effects of power," *American Economic Review*, 103(4): 1325–59.

Fehr, Ernst and Klaus M. Schmidt (1999). "A theory of fairness, competition and cooperation," *Quarterly Journal of Economics*, 114(3): 817–868.

Fong, Christina (2001). "Social preferences, self-interest, and the demand for redistribution," *Journal of Public Economics*, 82(2): 225–246.

Freundt, Jana, Holger Herz, and Leander Kopp (2023). "Intrinsic preferences for choice autonomy," CESifo Working Paper No. 10342.

Gangadharan, Lata, Philip J. Grossman, Kristy Jones, and C. Matthew Leister (2018). "Paternalistic giving: Restricting recipient choice," *Journal of Economic Behavior and Organization*, 151: 143–170.

Hausman, Daniel and Michael McPherson (2009). "Preference satisfaction and welfare economics," *Economics and Philosophy*, 25(1): 1–25.

Jacobsson, Fredric, Magnus Johannesson, and Lars Borgquist (2007). "Is altruism paternalistic?" *Economic Journal*, 117(520): 761–781.

Jung, Janice Y and Barbara A Mellers (2016). "American attitudes toward nudges." *Judgment & Decision Making*, 11(1).

Kaplow, Louis and Steven Shavell (2000). "Fairness versus welfare," *Harvard Law Review*, 114(961).

Kaplow, Louis and Steven Shavell (2001). "Any non-welfarist method of policy assessment violates the pareto principle," *Journal of Political Economy*, 109(2): 281–286.

Kaplow, Louis and Steven Shavell (2006). *Fairness versus welfare*, Cambridge, Massachusetts: Harvard University Press.

Kiessling, Lucas, Shyamal Chowdhury, Hannah Schildberg-Hörisch, and Matthias Sutter (2021). "Parental paternalism and patience," IZA DP No. 14030.

Konow, James (2000). "Fair shares: Accountability and cognitive dissonance in allocation decisions," *American Economic Review*, 90(4): 1072–1091.

Kuziemko, Ilyana, Michael I. Norton, Emmanuel Saez, and Stefanie Stantcheva (2015). "How elastic are preferences for redistribution? Evidence from randomized survey experiments," *American Economic Review*, 105(4): 1478–1508.

Le Grand, Julian and Bill New (2015). *Government Paternalism: Nanny State Or Helpful Friend?*, Princeton, NJ: Princeton University Press.

List, John A., Matthias Rodemeier, Sutanuka Roy, and Gregory K. Sun (2023). "Judging nudging: Understanding the welfare effects of nudges versus taxes," Working Paper 311520, National Bureau of Economic Research.

Nussbaum, Martha (2001). "Symposium on amartya sen's philosophy: Adaptive preferences and women's options," *Economics and Philosophy*, 17: 67–88.

OECD (2017). "Behavioural insights and public policy: Lessons from around the world," Technical report, OECD Publishing.

Pikulina, Elena S and Chloe Tergiman (2020). "Preferences for power," *Journal of Public Economics*, 185: 104173.

Reisch, Lucia A and Cass R Sunstein (2016). "Do europeans like nudges?" *Judgment and Decision making*, 11(4): 310–325.

Romano, Joseph P. and Michael Wolf (2005). "Stepwise multiple testing as formalized data snooping," *Econometrica*, 73(4): 1237–1282.

Romano, Joseph P. and Michael Wolf (2016). "Efficient computation of adjusted p-values for resampling-based stepdown multiple testing," *Statistics & Probability Letters*, 113(1): 38–40.

Sen, Amartya (2002). *Rationality and Freedom*, Harvard University Press.

Sunstein, Cass R., Lucia A. Reisch, and Julius Rauber (2018). "A worldwide consensus on nudging? not quite, but almost," *Regulation & Governance*, 12(1): 3–22.

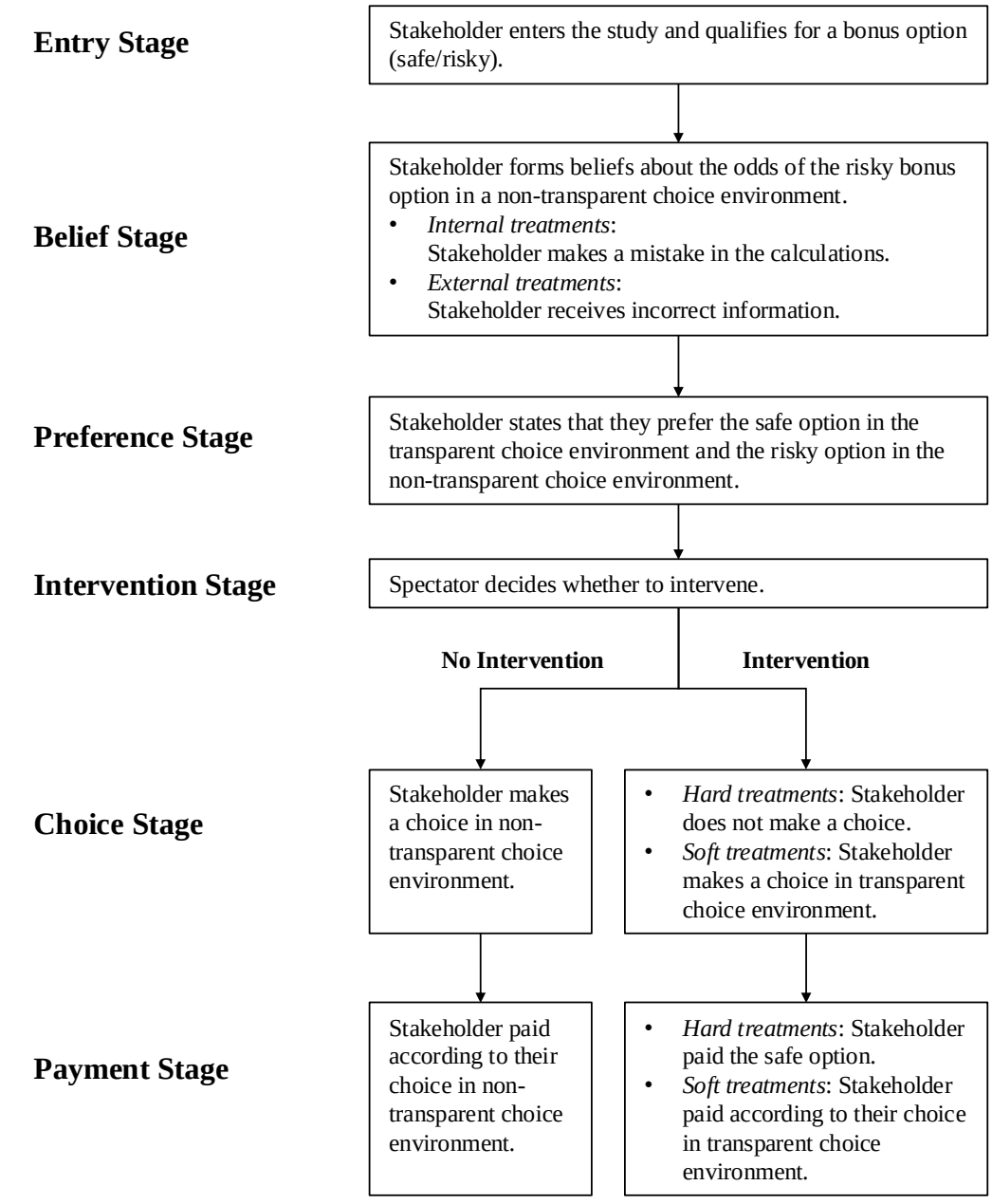
Sunstein, Cass R. and Richard H. Thaler (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*, New Haven: Yale University Press.

Tannenbaum, David, Craig R. Fox, and Todd Rogers (2017). "On the misplaced politics of behavioral policy interventions," *Nature Human Behaviour*, 1(0130): 1–7.

Thaler, Richard H. and Cass R. Sunstein (2003). "Libertarian paternalism," *American Economic Review*, 93(2): 175–179.

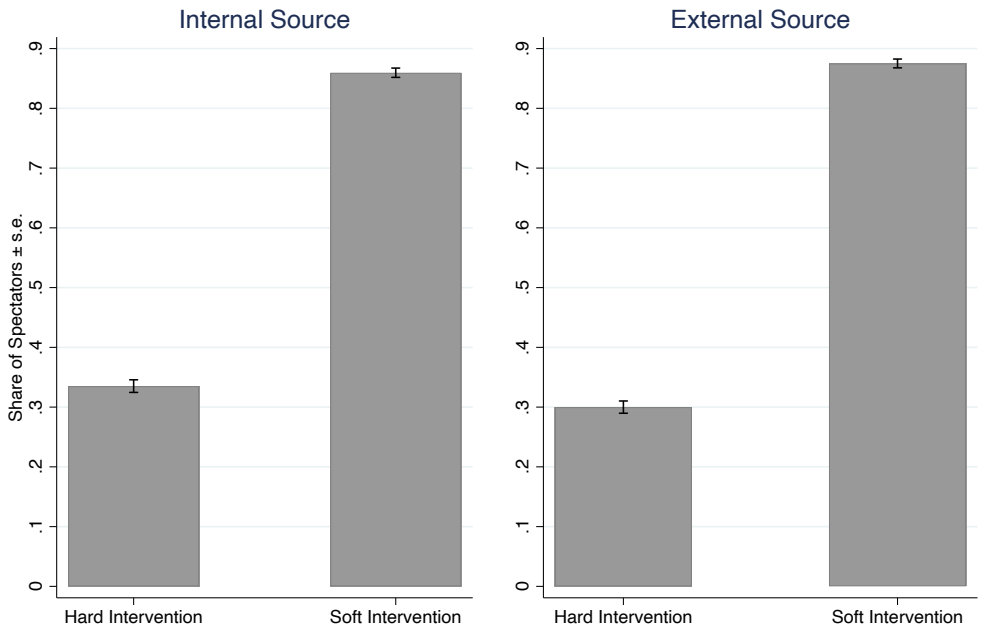
US Census Bureau (2018). "Data from www.census.gov, 2018–2020," Library Catalog: www.census.gov Section: Government.

Figure 1: Sequence of Events in Study 1



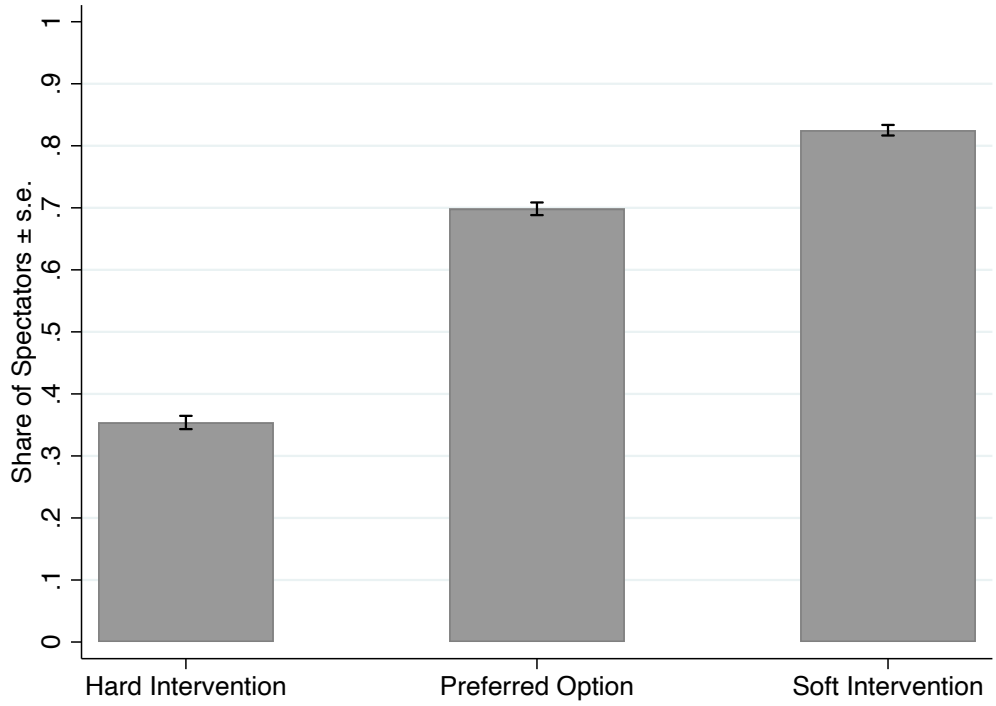
Notes: The figure shows the events for a stakeholder that was matched with a spectator. Stakeholders that stated a different preference ranking in the *Preference Stage* were not matched to a spectator.

Figure 2: Spectator Decisions by Treatment — Study 1



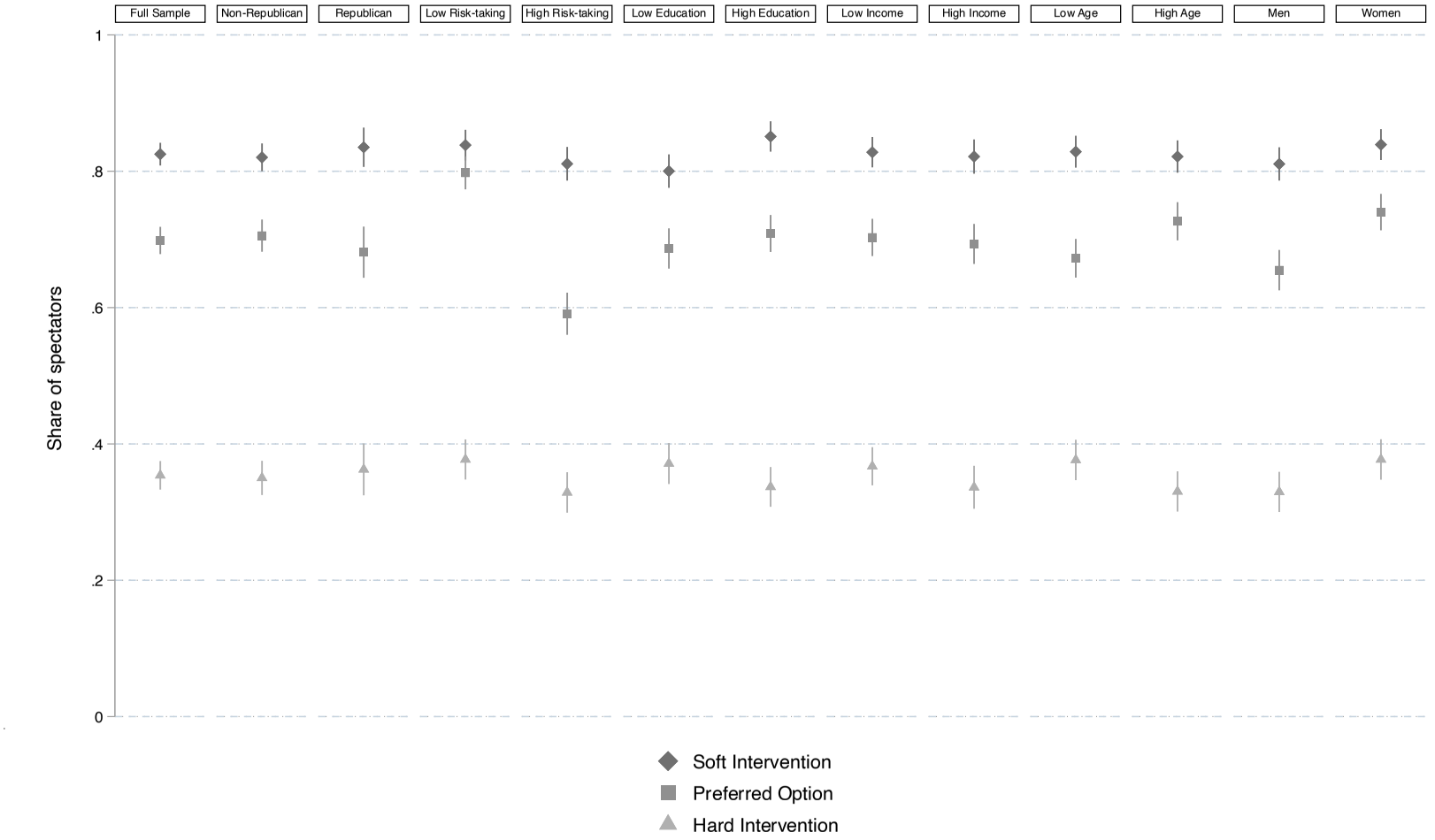
Note: The figure shows the share of spectators that intervene by treatment. The left panel shows the share of spectators intervening in treatments *Hard*×*Internal* and *Soft*×*Internal*. The right panel shows the share of spectators intervening in treatments *Hard*×*External* and *Soft*×*External*. The black bars indicate standard errors.

Figure 3: Spectator Decisions by Treatment — Study 2



Note: The left bar shows the share of spectators intervening in treatment *Hard*. The middle bar shows the share of spectators allocating the preferred safe option to the stakeholder in treatment *Welfare*. The right bar shows the share of spectators intervening in treatment *Soft*. The black bars indicate standard errors.

Figure 4: Spectator Decisions by Subgroup — Study 2



Note: The figure shows, for the full sample and for different subgroups in the general population, the share of spectators that intervene in treatment *Hard*, the share of spectators allocating the preferred safe option to the stakeholder in treatment *Welfare*, and the share of spectators that intervene in treatment *Soft*. Error bars indicate 95% confidence intervals. See Table A12 for the underlying regression results.

Table 1: Regression Results — Study 1

	(1)	(2)	(3)	(4)	(5)	(6)
Soft Intervention	0.550***	0.550***	0.524***	0.524***		
	(0.009)	(0.009)	(0.013)	(0.013)		
External Source			-0.035**	-0.035**	-0.009	-0.008
			(0.015)	(0.015)	(0.011)	(0.011)
Soft × External			0.051***	0.051***		
			(0.018)	(0.018)		
Republican		0.004		0.003		0.009
		(0.010)		(0.010)		(0.012)
High Risk Taking		-0.031***		-0.031***		-0.036***
		(0.009)		(0.009)		(0.011)
High Education		0.016		0.016		0.020*
		(0.010)		(0.010)		(0.012)
High Income		-0.026**		-0.026***		-0.025**
		(0.010)		(0.010)		(0.012)
High Age		-0.003		-0.003		-0.008
		(0.010)		(0.010)		(0.012)
Female		0.017*		0.016*		0.009
		(0.010)		(0.010)		(0.011)
Constant	0.318***	0.329***	0.335***	0.347***	0.596***	0.612***
	(0.007)	(0.013)	(0.011)	(0.015)	(0.008)	(0.015)
Observations	8,004	8,004	8,004	8,004	8,004	8,004
R^2	0.313	0.315	0.314	0.316	0.000	0.003

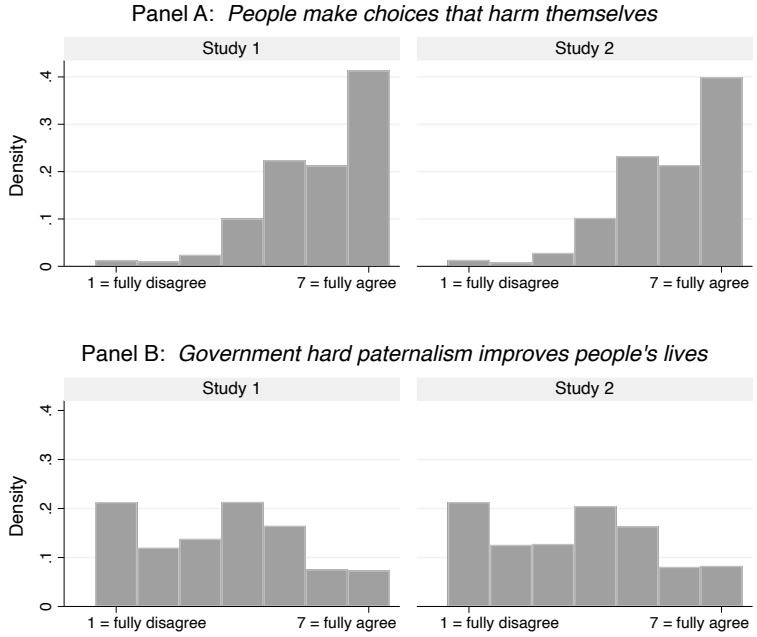
Notes: The table reports OLS regressions. The dependent variable is an indicator for whether the spectator chooses to intervene. The data from all four treatments is included. "Soft Intervention" is an indicator for the spectator being in a treatment with the soft intervention, "External Source" is an indicator for the spectator being in a treatment where the source of mistake is external. "Soft×External" is the interaction between these two variables. "Republican" is an indicator for identifying with the Republican party. "High Risk Taking," "High Education," "High Income," and "High Age" are indicator variables for having above-median willingness to take risks, education, income, and age, respectively. "Female" is an indicator for being female. The results are robust to adjusting for multiple-hypothesis testing and to using Probit models (see Online Appendix Tables A4 and A5). Robust standard errors in parentheses. * $p<.10$, ** $p<.05$, *** $p<.01$

Online Appendix

A Additional Figures and Tables

A.1 Additional Figures

Figure A1: Relevance of Experimental Context



Notes: The figure shows the distribution of agreement with the following statements: *People make choices that harm themselves* refers to the statement: "People make choices that harm their own well-being" (Panel A) and *Government hard paternalism improves people's lives* refers to the statement "The government can sometimes improve its citizens' well-being by restricting their freedom of choice" (Panel B). Spectators provided answers on a scale ranging from 1 = "fully disagree" to 7 = "fully agree." $n = 8,004$ for Study 1 and $n = 6,033$ for Study 2. We define "disagreement" with a statement as selecting a response smaller than the middle option 4. There is a strong positive association between abstaining from implementing the hard intervention in the experiment and disagreement with the view that the government can improve people's lives by means of hard paternalism. In a regression model where the dependent variable is the level of agreement on the *Government*-question (Panel B) and the independent variable is an indicator variable for whether the spectator chooses to intervene in a treatment with a hard intervention, the estimated coefficient is 0.298 ($p < 0.001$). The regression is estimated for all spectators in a treatment with a hard intervention, pooled for Study 1 and Study 2 ($n = 6,014$).

A.2 Sample Descriptives and Balancing Tables

Table A1: Sample Descriptives

	Study 1		Study 2		U.S. Population
	Mean	SD	Mean	SD	Mean
Female	0.51	(0.50)	0.51	(0.50)	0.51
Age 18–34	0.30	(0.46)	0.30	(0.46)	0.31
Age 35–44	0.18	(0.39)	0.17	(0.38)	0.18
Age 45–54	0.18	(0.38)	0.19	(0.39)	0.19
Age 55–64	0.16	(0.37)	0.16	(0.37)	0.16
Age 65–	0.18	(0.38)	0.18	(0.38)	0.17
Edu: Highschool or less	0.31	(0.46)	0.27	(0.44)	0.37
Edu: Some College	0.21	(0.41)	0.23	(0.42)	0.20
Edu: Bachelor or Associate	0.33	(0.47)	0.39	(0.49)	0.30
Edu: Master or above	0.15	(0.36)	0.12	(0.33)	0.14
Income $<$ 30,000	0.26	(0.44)	0.26	(0.44)	0.25
Income 30–60,000	0.26	(0.44)	0.30	(0.46)	0.25
Income 60–100,000	0.23	(0.42)	0.24	(0.43)	0.22
Income 100–150,000	0.14	(0.35)	0.13	(0.34)	0.14
Income $>$ 150,000	0.11	(0.32)	0.08	(0.27)	0.14
Republican	0.29	(0.45)	0.31	(0.46)	
Democrat	0.33	(0.47)	0.35	(0.48)	
Independent/Third Party	0.28	(0.45)	0.27	(0.44)	
Risk Taking: low (0–4)	0.23	(0.42)	0.22	(0.41)	
Risk Taking: median (5)	0.13	(0.34)	0.12	(0.33)	
Risk Taking: high (6–10)	0.64	(0.48)	0.66	(0.47)	
Observations	8,004		6,033		

Notes: Sample descriptives for spectators in Study 1 and Study 2. We asked the spectators to identify as either male or female, and we elicited the exact year of age. Education was elicited using the categories Less than High School, High School/GED, Some College, Associate's Degree, Bachelor's Degree, Master's Degree, Professional Degree (JD, MD), and Doctoral Degree. Income was elicited using the income brackets as shown in the table. The spectators were asked to identify as either "Republican," "Democrat," or "Independent/Third Party," and they had the option not to answer this question. "Risk Taking" was measured on a scale from 0 to 10, with 0 indicating "Completely unwilling to take risks" and 10 indicating "Very willing to take risks." We benchmark our sample composition against values for the population in the United States taken from the U.S. Census Bureau (2018).

Table A2: Balance Table Study 1

	Hard × Internal (1)	Soft × Internal (2)	Hard × External (3)	Soft × External (4)
Republican	0.003	0.003	-0.013	0.007
	(0.011)	(0.011)	(0.011)	(0.011)
High Risk Taking	0.007	-0.018*	0.003	0.008
	(0.010)	(0.010)	(0.010)	(0.010)
High Education	-0.002	0.004	-0.006	0.004
	(0.010)	(0.010)	(0.010)	(0.010)
High Income	0.006	0.007	-0.007	-0.006
	(0.011)	(0.011)	(0.011)	(0.011)
High Age	0.011	0.002	-0.002	-0.011
	(0.010)	(0.010)	(0.010)	(0.010)
Female	0.020**	-0.000	-0.006	-0.014
	(0.010)	(0.010)	(0.010)	(0.010)
p-value F-test	.559	.597	.768	. 621
Observations	8,004	8,004	8,004	8,004

Notes: The table reports regressions where the dependent variable is an indicator for being in the respective treatment (Hard × Internal, Soft × Internal, Hard × External, Soft × External). "Republican" is an indicator for identifying with the Republican party. "High Risk Taking," "High Education," "High Income," and "High Age" are indicator variables for having above-median willingness to take risks, education, income, and age, respectively. "Female" is an indicator for being female. Separate t-tests for differences across treatments confirm the findings. * $p<.10$, ** $p<.05$, *** $p<.01$

Table A3: Balance Table Study 2

	Hard (1)	Welfare (2)	Soft (3)
Republican	0.004	-0.022*	0.018
	(0.013)	(0.013)	(0.013)
High Risk Taking	-0.000	-0.005	0.005
	(0.012)	(0.013)	(0.012)
High Education	0.003	0.028**	-0.031**
	(0.013)	(0.013)	(0.013)
High Income	-0.025*	0.023*	0.002
	(0.013)	(0.013)	(0.013)
High Age	-0.001	-0.017	0.018
	(0.013)	(0.013)	(0.013)
Female	-0.002	0.009	-0.006
	(0.013)	(0.013)	(0.013)
p-value F-test	.708	.028**	.139
Observations	6,033	6,033	6,033

Notes: The table reports regressions where the dependent variable is an indicator for being in the respective treatment (Hard, Welfare, Soft). "Republican" is an indicator for identifying with the Republican party. "High Risk Taking," "High Education," "High Income," and "High Age" are indicator variables for having above-median willingness to take risks, education, income, and age, respectively. "Female" is an indicator for being female. Separate t-tests for differences across treatments confirm the findings. * $p<.10$, ** $p<.05$, *** $p<.01$

A.3 Additional Tables for Study 1

Table A4: MHT Corrections for Table 1

	(1)	(2)	(3)	(4)	(5)	(6)
Soft Intervention	0.550***	0.550***	0.524***	0.524***		
	[.001]	[.001]	[.001]	[.001]		
External Source			-0.035**	-0.035**	-0.009	-0.008
			[.026]	[.027]	[.395]	[.421]
Soft × External			0.051**	0.051**		
			[.013]	[.011]		
Republican		0.004		0.003		0.009
High Risk Taking		-0.031***		-0.031***		-0.036***
High Education		0.016		0.016		0.020*
High Income		-0.026**		-0.026***		-0.025**
High Age		-0.003		-0.003		-0.008
Female		0.017*		0.016*		0.009
Constant	0.318***	0.329***	0.335***	0.347***	0.596***	0.612***
Observations	8,004	8,004	8,004	8,004	8,004	8,004
R^2	0.313	0.315	0.314	0.316	0.000	0.003

Notes: The table reports OLS regressions. The dependent variable is an indicator for whether the spectator chooses to intervene. The treatment with the hard intervention and the internal source of mistake serves as omitted category. "Soft Intervention" is an indicator for the spectator being in a treatment with a soft intervention, "External Source" is an indicator for the spectator being in a treatment where the source of mistake is external. "Soft×External" is the interaction between these two variables. "Republican" is an indicator for identifying with the Republican party. "High Risk Taking," "High Education," "High Income," and "High Age" are indicator variables for having above-median willingness to take risks, education, income, and age, respectively. "Female" is an indicator for being female. The p-values are adjusted for multiple hypothesis testing (MHT), using the Romano-Wolf stepdown procedure (Romano and Wolf, 2005, 2016). We correct for multiple treatments within (i) Columns (1), (3), and (5), and (ii) within Columns (2), (4), and (6), respectively. Adjusted p-values are reported in brackets. * $p<.10$, ** $p<.05$, *** $p<.01$

Table A5: Probit Models for Table 1

	(1)	(2)	(3)	(4)	(5)	(6)
Soft Intervention	1.588***	1.594***	1.504***	1.509***		
	(0.032)	(0.033)	(0.045)	(0.045)		
External Source			-0.099**	-0.099**	-0.023	-0.021
			(0.041)	(0.041)	(0.028)	(0.028)
Soft × External			0.171***	0.174***		
			(0.065)	(0.065)		
Republican		0.012		0.011		0.023
		(0.036)		(0.036)		(0.032)
High Risk Taking		-0.108***		-0.109***		-0.094***
		(0.033)		(0.033)		(0.029)
High Education		0.063*		0.062*		0.053*
		(0.034)		(0.034)		(0.030)
High Income		-0.090**		-0.090**		-0.065**
		(0.035)		(0.035)		(0.031)
High Age		-0.009		-0.010		-0.021
		(0.034)		(0.034)		(0.030)
Female		0.063*		0.062*		0.023
		(0.033)		(0.033)		(0.030)
Constant	-0.474***	-0.444***	-0.426***	-0.393***	0.244***	0.286***
	(0.021)	(0.043)	(0.029)	(0.048)	(0.020)	(0.039)
Observations	8,004	8,004	8,004	8,004	8,004	8,004

Notes: The table reports probit models corresponding to the OLS models shown in Table 1. Robust standard errors in parentheses. * $p<.10$, ** $p<.05$, *** $p<.01$

Table A6: Heterogeneity: Nature of Intervention

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Soft Intervention	0.552***	0.539***	0.521***	0.531***	0.535***	0.545***	0.477***
	(0.011)	(0.012)	(0.013)	(0.013)	(0.013)	(0.013)	(0.024)
Soft × Republican	-0.009						-0.015
	(0.020)						(0.020)
Republican	0.004						0.010
	(0.016)						(0.016)
Soft × High Risk Taking		0.023					0.029
		(0.018)					(0.019)
High Risk Taking		-0.045***					-0.046***
		(0.015)					(0.015)
Soft × High Education			0.060***				0.053***
			(0.018)				(0.020)
High Education			-0.025*				-0.011
			(0.015)				(0.016)
Soft × High Income				0.039**			0.023
				(0.018)			(0.020)
High Income				-0.046***			-0.037**
				(0.015)			(0.016)
Soft × High Age					0.030		0.028
					(0.018)		(0.019)
High Age					-0.016		-0.017
					(0.015)		(0.015)
Soft × Female						0.010	0.027
						(0.018)	(0.019)
Female						0.019	0.004
						(0.015)	(0.015)
Constant	0.317***	0.339***	0.329***	0.339***	0.326***	0.308***	0.365***
	(0.009)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.019)
Soft +	0.543***	0.562***	0.581***	0.570***	0.565***	0.555***	
Soft × Indicator	(0.017)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	
Observations	8,004	8,004	8,004	8,004	8,004	8,004	8,004
R^2	0.313	0.314	0.314	0.314	0.313	0.313	0.317

Notes: The table reports OLS regressions. The dependent variable is an indicator for whether the spectator chooses to intervene. Treatments *Hard*×*Internal* and *Hard*×*External* serve as omitted category. "Soft Intervention" is an indicator for the spectator being in treatment *Soft*×*Internal* or *Soft*×*External*. "Soft×..." denotes the interaction between "Soft Intervention" and the following indicator variables. "Republican" is an indicator for identifying with the Republican party. "High Risk Taking," "High Education," "High Income," and "High Age" are indicator variables for having above-median willingness to take risks, education, income, and age, respectively. "Female" is an indicator for being female. In models (1) to (6) we control for the respective non-interacted indicator variable and in model (7) we control for all six non-interacted indicator variables. The results are robust to adjusting for multiple-hypothesis testing (see Table A7). Robust standard errors in parentheses. * $p<.10$, ** $p<.05$, *** $p<.01$

Table A7: MHT Corrections for Table A6

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Soft Intervention	0.552***	0.539***	0.521***	0.531***	0.535***	0.545***	0.477***
	[.001]	[.001]	[.001]	[.001]	[.001]	[.001]	[.001]
Soft × Republican	-0.009						-0.015
	[.827]						[.466]
Soft × High Risk Taking		0.023					0.029
		[.482]					[.436]
Soft × High Education			0.060***				0.053**
			[.005]				[.038]
Soft × High Income				0.039			0.023
				[.145]			[.436]
Soft × High Age					0.030		0.028
					[.331]		[.436]
Soft × Female						0.010	0.027
						[.827]	[.436]
Observations	8,004	8,004	8,004	8,004	8,004	8,004	8,004
R^2	0.313	0.314	0.314	0.314	0.313	0.313	0.317

Notes: The table reports OLS regressions. The dependent variable is an indicator for whether the spectator chooses to intervene. Treatments *Hard*×*Internal* and *Hard*×*External* serve as omitted category. "Soft Intervention" is an indicator for the spectator being in treatment *Soft*×*Internal* or *Soft*×*External*. "Soft×..." denotes the interaction between "Soft Intervention" and the following indicator variables. "Republican" is an indicator for identifying with the Republican party. "High Risk Taking," "High Education," "High Income," and "High Age" are indicator variables for having above-median willingness to take risks, education, income, and age, respectively. "Female" is an indicator for being female. In models (1) to (6) we control for the respective non-interacted indicator variable and in model (7) we control for all six non-interacted indicator variables. The p-values are adjusted for multiple hypothesis testing (MHT), using the Romano-Wolf stepdown procedure (Romano and Wolf, 2005, 2016). We correct for multiple subgroup comparisons within Columns (1)–(6) and within Column (7), respectively. Adjusted p-values are reported in brackets. * $p<.10$, ** $p<.05$, *** $p<.01$

Table A8: Heterogeneity: Source of Mistake

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
External Source	-0.018	-0.020	-0.011	-0.004	0.005	-0.004	-0.007
	(0.013)	(0.015)	(0.015)	(0.015)	(0.015)	(0.016)	(0.028)
Ext × Republican	0.032						0.037
	(0.024)						(0.025)
Republican	-0.011						-0.009
	(0.017)						(0.017)
Ext × High Risk Taking		0.026					0.021
		(0.022)					(0.023)
High Risk Taking		-0.051***					-0.047***
		(0.016)					(0.016)
Ext × Education			0.005				0.016
			(0.022)				(0.023)
High Education			0.008				0.013
			(0.016)				(0.016)
Ext × Income				-0.010			-0.018
				(0.022)			(0.024)
High Income				-0.018			-0.016
				(0.016)			(0.017)
Ext × Age					-0.028		-0.031
					(0.022)		(0.023)
High Age					0.011		0.007
					(0.016)		(0.016)
Ext × Female						-0.010	-0.011
						(0.022)	(0.023)
Female						0.022	0.014
						(0.016)	(0.016)
Constant	0.600***	0.619***	0.593***	0.605***	0.591***	0.585***	0.611***
	(0.009)	(0.010)	(0.011)	(0.011)	(0.011)	(0.011)	(0.020)
External Source + Ext × Indicator	0.014	0.006	-0.006	-0.014	-0.023	-0.013	
	(0.020)	(0.016)	(0.016)	(0.016)	(0.016)	(0.015)	
Observations	8,004	8,004	8,004	8,004	8,004	8,004	8,004
R^2	0.000	0.002	0.000	0.001	0.000	0.000	0.003

Notes: The table reports OLS regressions. The dependent variable is an indicator for whether the spectator chooses to intervene. Treatments *Hard*×*Internal* and *Soft*×*Internal* serve as omitted category. "External Source" is an indicator for the spectator being in treatment *Hard*×*External* or *Soft*×*External*. "Ext×..." denotes the interaction between "External Source" and the following indicator variables. "Republican" is an indicator for identifying with the Republican party. "High Risk Taking," "High Education," "High Income," and "High Age" are indicator variables for having above-median willingness to take risks, education, income, and age, respectively. "Female" is an indicator for being female. In models (1) to (6) we control for the respective non-interacted indicator variable and in model (7) we control for all six non-interacted indicator variables. The results are robust to adjusting for multiple-hypothesis testing (see Table A9). Robust standard errors in parentheses. * $p<.10$, ** $p<.05$, *** $p<.01$

Table A9: MHT Corrections for Table A8

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
External Source	-0.018	-0.020	-0.011	-0.004	0.005	-0.004	-0.007
	[.745]	[.749]	[.949]	[.995]	[.995]	[.995]	[.900]
Ext × Republican	0.032						0.037
	[.760]						[.579]
Ext × High Risk Taking		0.026					0.021
		[.783]					[.837]
Ext × Education			0.005				0.016
			[.995]				[.900]
Ext × Income				-0.010			-0.018
				[.991]			[.900]
Ext × Age					-0.028		-0.031
					[.760]		[.623]
Ext × Female						-0.010	-0.011
						[.991]	[.900]
Observations	8,004	8,004	8,004	8,004	8,004	8,004	8,004
R^2	0.000	0.002	0.000	0.001	0.000	0.000	0.003

Notes: The table reports OLS regressions. The dependent variable is an indicator for whether the spectator chooses to intervene. Treatments *Hard*×*Internal* and *Soft*×*Internal* serve as omitted category. "External Source" is an indicator for the spectator being in treatment *Hard*×*External* or *Soft*×*External*. "Ext×..." denotes the interaction between "External Source" and the following indicator variables. "Republican" is an indicator for identifying with the Republican party. "High Risk Taking," "High Education," "High Income," and "High Age" are indicator variables for having above-median willingness to take risks, education, income, and age, respectively. "Female" is an indicator for being female. In models (1) to (6) we control for the respective non-interacted indicator variable and in model (7) we control for all six non-interacted indicator variables. The p-values are adjusted for multiple hypothesis testing (MHT), using the Romano-Wolf stepdown procedure (Romano and Wolf, 2005, 2016). We correct for multiple subgroup comparisons within Columns (1)–(6) and within Column (7), respectively. Adjusted p-values are reported in brackets. * $p<.10$, ** $p<.05$, *** $p<.01$

Table A10: Heterogeneity: Political Orientation

	Full Study 1 Sample			Only Republicans and Democrats		
	(1) Republicans	(2) Non-Republicans	(3) Fully interacted	(4) Republicans	(5) Democrats	(6) Fully interacted
Soft Intervention	0.534***	0.520***	0.520***	0.534***	0.520***	0.520***
	(0.024)	(0.016)	(0.016)	(0.024)	(0.023)	(0.023)
External Source	-0.004	-0.048***	-0.048***	-0.004	-0.048*	-0.048*
	(0.028)	(0.017)	(0.017)	(0.028)	(0.026)	(0.026)
Soft × External	0.018	0.064***	0.064***	0.018	0.058*	0.058*
	(0.034)	(0.022)	(0.022)	(0.034)	(0.031)	(0.031)
Republican			-0.019			-0.031
			(0.023)			(0.027)
Soft × Republican			0.014			0.014
			(0.029)			(0.033)
External × Republican			0.044			0.044
			(0.033)			(0.038)
Soft × Ext × Republican			-0.045			-0.039
			(0.040)			(0.046)
Constant	0.322***	0.341***	0.341***	0.322***	0.353***	0.353***
	(0.019)	(0.013)	(0.013)	(0.019)	(0.019)	(0.019)
Observations	2,316	5,688	8,004	2,316	2,658	4,974
R^2	0.307	0.317	0.314	0.307	0.316	0.312

Notes: The table reports OLS regressions. The dependent variable is an indicator for whether the spectator chooses to intervene. Treatment *Hard*×*Internal* serves as omitted category. "Soft Intervention" is an indicator for the spectator being in treatment *Soft*×*Internal* or *Soft*×*External*. "External Source" is an indicator for the spectator being in treatment *Hard*×*External* or *Soft*×*External*. "Republican" is an indicator for identifying with the Republican party. "...×..." denotes the respective interaction terms. Columns (1) and (2) are estimated separately for the sub-samples of Republicans and non-Republicans. Column (3) is estimated for the full sample. Column (4) is identical to Column (1). Column (5) is estimated separately for the sub-samples of Democrats. Column (6) is estimated for the sub-sample of Republicans and Democrats (excluding participants who self-identify as "Independent/Third Party" or did not report a political affiliation). Robust standard errors in parentheses. * $p<.10$, ** $p<.05$, *** $p<.01$

A.4 Additional Tables for Study 2

Table A11: Results for Study 2 — OLS, MHT Corrections, and Probit Models

	Main		MHT		Probit	
	(1)	(2)	(3)	(4)	(5)	(6)
Soft Intervention	0.127***	0.128***	0.127***	0.128***	0.415***	0.421***
	(0.013)	(0.013)	[.001]	[.001]	(0.044)	(0.044)
Hard Intervention	-0.344***	-0.344***	-0.344***	-0.344***	-0.895***	-0.908***
	(0.015)	(0.015)	[.001]	[.001]	(0.041)	(0.041)
Republican		0.012		0.012		0.037
		(0.012)		(0.012)		(0.039)
High Risk Taking		-0.094***		-0.094***		-0.289***
		(0.012)		(0.012)		(0.036)
High Education		0.030**		0.030**		0.099**
		(0.012)		(0.012)		(0.038)
High Income		-0.009		-0.009		-0.030
		(0.013)		(0.013)		(0.039)
High Age		-0.015		-0.015		-0.046
		(0.012)		(0.012)		(0.037)
Female		0.050***		0.050***		0.156***
		(0.012)		(0.012)		(0.037)
Constant	0.698***	0.711***	0.698***	0.711***	0.520***	0.563***
	(0.010)	(0.017)	(0.010)	(0.017)	(0.029)	(0.052)
Observations	6033	6033	6033	6033	6033	6033
R^2	0.169	0.182	0.169	0.182		

Notes: The table reports OLS regressions in Columns (1)–(4) and probit models in Columns (5)–(6). The dependent variable is an indicator for whether the spectator chooses to intervene or select the safe option. Treatment *Welfare* serves as omitted category. "Soft Intervention" and "Hard Intervention" are indicator variables for the spectator being in treatment *Soft* and *Hard*, respectively. "Republican" is an indicator for identifying with the Republican party. "High Risk Taking," "High Education," "High Income," and "High Age" are indicator variables for having above-median willingness to take risks, education, income, and age, respectively. "Female" is an indicator for being female. Robust standard errors in parentheses. The p-values in Columns (3) and (4) are adjusted for multiple hypothesis testing (MHT), using the Romano-Wolf stepdown procedure (Romano and Wolf, 2005, 2016). We correct for multiple treatments within Column (1) and within Column (2), respectively. Adjusted p-values are reported in brackets. * $p<.10$, ** $p<.05$, *** $p<.01$

Table A12: Heterogeneity in Study 2

	Republican (1)	Risk Taking (2)	Education (3)	Income (4)	Age (5)	Female (6)
Soft Intervention	0.115***	0.041**	0.113***	0.125***	0.156***	0.156***
	(0.016)	(0.017)	(0.020)	(0.018)	(0.019)	(0.020)
Hard Intervention	-0.355***	-0.421***	-0.315***	-0.336***	-0.296***	-0.325***
	(0.018)	(0.019)	(0.021)	(0.020)	(0.021)	(0.021)
Soft × Indicator	0.039	0.179***	0.029	0.003	-0.061**	-0.057**
	(0.029)	(0.026)	(0.027)	(0.027)	(0.027)	(0.027)
Hard × Indicator	0.037	0.158***	-0.056*	-0.021	-0.100***	-0.038
	(0.032)	(0.029)	(0.030)	(0.030)	(0.029)	(0.029)
Indicator	-0.024	-0.207***	0.022	-0.010	0.054***	0.085***
	(0.023)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)
Constant	0.706***	0.798***	0.687***	0.703***	0.672***	0.655***
	(0.012)	(0.012)	(0.015)	(0.014)	(0.014)	(0.015)
Observations	6033	6033	6033	6033	6033	6033
R^2	0.169	0.185	0.170	0.169	0.171	0.172

Notes: The table reports OLS regressions. The dependent variable is an indicator for whether the spectator chooses to intervene. Treatment *Welfare* serves as omitted category. "Soft Intervention" is an indicator for the spectator being in treatment *Soft*. "Hard Intervention" is an indicator for the spectator being in treatment *Hard*. "Soft × Indicator" ("Hard × Indicator") denotes the interaction between "Soft Intervention" ("Hard Intervention") and the following indicator variables. Column (1): "Republican" is an indicator for identifying with the Republican party. Column (2)–(5): "Risk Taking," "Education," "Income," and "Age" are indicator variables for having above-median willingness to take risks, education, income, and age, respectively. Column (6): "Female" is an indicator for being female. Robust standard errors in parentheses. * $p<.10$, ** $p<.05$, *** $p<.01$

B Experimental Procedures

In Section B.1, we describe how we elicited the stakeholders' preferences over the bonus options and how we matched them to spectators. In Section B.2, we provide the experimental instructions for the spectators.

B.1 Stakeholders

Here we provide the details on how we elicited the preferences of the stakeholders recruited on the online labor platform (MTurk) and the matching protocol.

Preference Elicitation. We elicited the stakeholders' preferences over the safe and the risky bonus option in both the transparent choice environment and in one of two conditions of the non-transparent choice environment.

In the transparent choice environment, all stakeholders received the following instructions:

At the end of this study, you can receive a bonus payment (depending on your choices). Suppose you could choose between the following two bonus options:

Safe option:	*A bonus of 4 USD for sure.*
Risky option:	*This option is a lottery. It pays a bonus of 10 USD or nothing, where the two outcomes are equally likely.*

Which of these two bonus options would you prefer?

◯ *Safe option*

◯ *Risky option*

In the non-transparent choice environment, the stakeholders who were randomized into *internal* condition received a signal that would allow a Bayesian individual to correctly calculate that the likelihood of receiving USD 10 is 50%. Stakeholders who fall prey to base-rate neglect, however, would infer that the likelihood of receiving USD 10 is higher than it actually is. The instructions in the *internal* condition are as follows:

Safe option:	*A bonus of 4 USD for sure*
Risky option:	*This option gives you a ticket for a lottery. You win a bonus of 10 USD, if you have a winning ticket. A random ticket wins with a probability of 1%. However, your ticket was pre-tested and according to the pre-test it is a winning ticket. The pre-test correctly identifies winning and losing tickets in 99% of the cases.*

Which of these two bonus options would you prefer?

◯ *Safe option*

◯ *Risky option*

Stakeholders who were randomized into *external* condition received a signal about the likelihood of receiving USD 10 and were truthfully informed that the signal is not always exactly precise, but that the average of all signals sent is correct. Some of these stakeholders received the incorrect signal that the likelihood of receiving USD 10 is 75% (while the true value is 50%). Stakeholders who naively follow the signal would infer that the likelihood of receiving USD 10 is higher than it actually is. To ensure that the average of all signals sent is correct (i.e., to ensure that participants are not deceived), we also implemented signals that the likelihood of receiving USD 10 is lower than it actually is. The instructions in the *external* condition for the 75%-signal are as follows:

Safe option:	*A bonus of 4 USD for sure*
Risky option:	*This option is a lottery. It pays a bonus of 10 USD with a certain probability and nothing otherwise. You are provided with a signal about the probability that the lottery pays the 10 USD (the signal is not always exactly precise; however, the average of all signals sent is correct). Your signal about the probability of getting 10 USD is 75%.*

Which of these two bonus options would you prefer?

◯ *Safe option*

◯ *Risky option*

Matching. Only stakeholders who prefer the safe option in the transparent choice environment but prefer the risky option in the non-transparent choice environment

were matched to a spectator. As pre-registered, stakeholders with a different preference profile (e.g., those who prefer the risky option in the transparent choice environment or those who prefer the safe option in the non-transparent choice environment because they received a signal that the likelihood of receiving USD 10 is lower than it actually is) were not matched to a spectator and simply received their payments for participation in the study.

A stakeholder who was assigned to the *internal* condition was matched to a spectator who was randomized into a treatment with internal source of mistake. Likewise, a stakeholder who was assigned to the *external* condition was matched to a spectator who was randomized into a treatment with external source of mistake. Stakeholders were then presented with the decision scenario resulting from the intervention choice of their matched spectator: If a spectator chose not to intervene, the matched stakeholder made their decision in the non-transparent choice environment, where they had indicated a preference for the risky option. If a spectator intervened, the matched stakeholder either received the safe option directly (hard intervention) or made their decision in the transparent choice environment, where they had indicated a preference for the safe option (soft intervention). Finally, stakeholders received the bonus that was either assigned to them (in the case of a hard intervention) or that they chose themselves (in the case of no intervention or soft intervention).

Based on the 5:1 matching ratio between spectators and stakeholders, we recruited stakeholders until reaching the required numbers: 1,601 stakeholders to match with the 8,004 spectators in Study 1 and 1,207 stakeholders to match with the 6,033 spectators in Study 2.

B.2 Spectators

Here we provide the instructions for the spectators in the four different treatments implemented in Study 1 and in the additional treatment implemented in Study 2. Bold text, underlining, tables, etc. appear as in the original screen.

B.2.1 Hard Intervention and Internal Source of Mistake (Study 1/Study 2)

We now ask you to make a decision that may have real consequences for another person (one out of five respondents to this survey are randomly selected and their choice will be implemented).

This other person was hired to do some work. After completing the work, the person was informed that he or she will get a bonus. There are two bonus options available:

Safe option:	*a bonus of 4 USD for sure*
Risky option:	*either a bonus of 10 USD or nothing, where the two outcomes are equally likely*

When the person was informed about the two options, the risky option was not presented as in the table above. Rather, the person had to calculate the likelihoods of the two outcomes of the risky option. The person made a mistake in the calculations.

*As a result, the person prefers **the risky option**. However, had the person calculated the likelihoods correctly, he or she would have preferred **the safe option**.*

The person has not yet made a choice. You can now decide between two alternatives:

- ◯ ***Restrict choice:** The person will not have the opportunity to make a choice and will receive the safe option.*
- ◯ ***Do not restrict choice:** The person will have the opportunity to make a choice between the safe and the risky option.*

The person will not be informed about your involvement.

B.2.2 Soft Intervention and Internal Source of Mistake (Study 1/Study 2)

We now ask you to make a decision that may have real consequences for another person (one out of five respondents to this survey are randomly selected and their choice will be implemented).

This other person was hired to do some work. After completing the work, the person was informed that he or she will get a bonus. There are two bonus options available:

Safe option:	*a bonus of 4 USD for sure*
Risky option:	*either a bonus of 10 USD or nothing, where the two outcomes are equally likely*

When the person was informed about the two options, the risky option was not presented as in the table above. Rather, the person had to calculate the likelihoods of the two outcomes of the risky option. The person made a mistake in the calculations.

*As a result, the person prefers **the risky option**. However, had the person calculated the likelihoods correctly, he or she would have preferred **the safe option**.*

The person has not yet made a choice. You can now decide between two alternatives:

- ◯ ***Provide information: The person will be informed about the correct likelihoods of the two outcomes in the risky option before he or she makes a choice between the safe and the risky option.***
- ◯ ***Do not provide information: The person will receive no additional information before he or she makes a choice between the safe and the risky option.***

The person will not be informed about your involvement.

B.2.3 Hard Intervention and External Source of Mistake (Study 1)

We now ask you to make a decision that may have real consequences for another person (one out of five respondents to this survey are randomly selected and their choice will be implemented).

This other person was hired to do some work. After completing the work, the person was informed that he or she will get a bonus. There are two bonus options available:

Safe option:	*a bonus of 4 USD for sure*
Risky option:	*either a bonus of 10 USD or nothing, where the two outcomes are equally likely*

When the person was informed about the two options, the risky option was not presented as in the table above. Rather, the person was unlucky and received incorrect information about the likelihoods of the two outcomes of the risky option.

*As a result, the person prefers **the risky option**. However, had the person received correct information about the likelihoods, he or she would have preferred **the safe option**.*

The person has not yet made a choice. You can now decide between two alternatives:

- ◯ ***Restrict choice:* The person will not have the opportunity to make a choice and will receive the safe option.**
- ◯ ***Do not restrict choice:* The person will have the opportunity to make a choice between the safe and the risky option.**

The person will not be informed about your involvement.

B.2.4 Soft Intervention and External Source of Mistake (Study 1)

We now ask you to make a decision that may have real consequences for another person (one out of five respondents to this survey are randomly selected and their choice will be implemented).

This other person was hired to do some work. After completing the work, the person was informed that he or she will get a bonus. There are two bonus options available:

Safe option:	*a bonus of 4 USD for sure*
Risky option:	*either a bonus of 10 USD or nothing, where the two outcomes are equally likely*

When the person was informed about the two options, the risky option was not presented as in the table above. Rather, the person was unlucky and received incorrect information about the likelihoods of the two outcomes of the risky option.

*As a result, the person prefers **the risky option**. However, had the person received correct information about the likelihoods, he or she would have preferred **the safe option**.*

The person has not yet made a choice. You can now decide between two alternatives:

- ◯ ***Provide information: The person will be informed about the correct likelihoods of the two outcomes in the risky option before he or she makes a choice between the safe and the risky option.***
- ◯ ***Do not provide information: The person will receive no additional information before he or she makes a choice between the safe and the risky option.***

The person will not be informed about your involvement.

B.2.5 Welfare and Internal Source of Mistake (Study 2)

We now ask you to make a decision that may have real consequences for another person (one out of five respondents to this survey are randomly selected and their choice will be implemented).

This other person was hired to do some work. After completing the work, the person was informed that he or she will get a bonus. There are two bonus options available:

Safe option:	*a bonus of 4 USD for sure*
Risky option:	*either a bonus of 10 USD or nothing, where the two outcomes are equally likely*

When the person was informed about the two options, the risky option was not presented as in the table above. Rather, the person had to calculate the likelihoods of the two outcomes of the risky option. The person made a mistake in the calculations.

*As a result, the person prefers **the risky option**. However, had the person calculated the likelihoods correctly, he or she would have preferred **the safe option**.*

The person has not yet made a choice. You can now decide between two alternatives:

- ◯ ***Restrict choice to safe option: The person will not have the opportunity to make a choice and will receive the safe option.***
- ◯ ***Restrict choice to risky option: The person will not have the opportunity to make a choice and will receive the risky option.***

The person will not be informed about your involvement.

C Pre-Analysis Plans

The pre-analysis plans were uploaded to the AEA Social Science Registry on August 28, 2019 (for Study 1), and on January 17, 2020 (for Study 2) and can be found here.

We closely follow the pre-analysis plans, with minor deviations:

1. We make semantic changes (changing the reference category, changing labels) to make the paper and the results easier to read.

2. We use a slightly smaller set of control variables than pre-specified (we do not control for region, marital status, and number of children), and we use indicator variables defined by median splits. We do this to simplify the presentation and interpretation of the results. Tables C1 and C2 show that the results shown in the main Tables 1 and A11 are unaffected by using the pre-specified controls.

3. We only specified the heterogeneity analysis in the pre-analysis plan for Study 1, with a focus on political orientation. In the paper, we report the heterogeneity analysis for both studies, and also with respect to willingness to take risks, education, income, age, and gender.

Table C1: Table 1 with Pre-Specified Controls

	(1)	(2)	(3)	(4)	(5)	(6)
Soft Intervention	0.550***	0.550***	0.524***	0.525***		
	(0.009)	(0.009)	(0.013)	(0.013)		
External Source			-0.035**	-0.035**	-0.008	-0.008
			(0.015)	(0.015)	(0.011)	(0.011)
Soft × External			0.051***	0.052***		
			(0.018)	(0.018)		
Constant	0.329***	0.376***	0.347***	0.394***	0.612***	0.611***
	(0.013)	(0.029)	(0.015)	(0.030)	(0.015)	(0.029)
Controls (Table 2)	Yes	No	Yes	No	Yes	No
Controls (Pre-plan)	No	Yes	No	Yes	No	Yes
Observations	8,004	8,004	8,004	8,004	8,004	8,004
R^2	0.315	0.316	0.316	0.317	0.003	0.003

Notes: The table reports OLS regressions. The dependent variable is an indicator for whether the spectator chooses to intervene. "Soft Intervention" is an indicator for the spectator being in a treatment with the soft intervention, "External Source" is an indicator for the spectator being in a treatment where the source of mistake is external. "Soft×External" is the interaction between these two variables. In models (1), (3), and (5), we include the set of controls as in Table 1. In models (2), (4), and (6), we include the set of controls as specified in the pre-analysis plan. Results are practically identical across specifications. Robust standard errors in parentheses. * $p<.10$, ** $p<.05$, *** $p<.01$

Table C2: Table A11 with Pre-Specified Controls

	(1)	(2)
Soft Intervention	0.128***	0.128***
	(0.013)	(0.013)
Hard Intervention	-0.344***	-0.345***
	(0.015)	(0.015)
Constant	0.711***	0.827***
	(0.017)	(0.036)
Controls (Table 3)	Yes	No
Controls (Pre-plan)	No	Yes
Observations	6,033	6,033
R^2	0.182	0.185

Notes: The table reports OLS regressions. The dependent variable is an indicator for whether the spectator chooses to intervene or select the safe option. Treatment *Welfare* serves as omitted category. "Soft Intervention" and "Hard Intervention" are indicators for the spectator being in treatment *Soft* and *Hard*, respectively. In model (1), we include the set of controls as in Table A11, Column (2). In model (2), we include the set of controls as specified in the pre-analysis plan. Results are practically identical across specifications. Robust standard errors in parentheses. * $p<.10$, ** $p<.05$, *** $p<.01$