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 (2023) : 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/10419/275656

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

May 2023

Free to Fail? Paternalistic Preferences in the United States*

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

May 16, 2023

Abstract

We study paternalistic preferences in two large-scale experiments with participants from the general population in the United States. Spectators decide whether to intervene to prevent a stakeholder, who is mistaken about the choice set, from making a choice that is not aligned with the stakeholders' own preferences. We find causal evidence for the nature of the intervention being of great importance for the spectators' willingness to intervene. Only a minority 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. This finding holds regardless of the stakeholder's responsibility for being mistaken about the choice set – whether the source of mistake is internal or external – and in different subgroups of the population. We introduce a theoretical framework with two paternalistic types – libertarian paternalists and welfarists – and show that the two types can account for most of the spectator behavior. We estimate that about half of the spectators are welfarists and that about a third are libertarian paternalists. Our results shed light on attitudes toward paternalistic policies and the broad support for soft interventions.

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

*We thank Sandro Ambuehl, Ceren Ay, Jill D. Berrick, Stefano DellaVigna, Martin Dufwenberg, Mathias Ekström, Dirk Engelmann, Christine Exley, Eleonora Freddi, Uri Gneezy, Hege S. Helland, Holger Herz, Ariel Kalil, Judd Kessler, Botond Kőszegi, Dorothea Kübler, Mathea Loen, Audun Løvlie, George Loewenstein, Matt Lowe, Muriel Niederle, Nick Netzer, Christopher Roth, Barbara Ruiken, Erik Ø. Sørensen, Krishna Srinivasan, Matthias Sutter, Georg Weizsäcker, Ludger Woessmann and numerous seminar and conference participants for very helpful suggestions and comments. Camilla Allocchio provided excellent research support. 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 was preregistered in the AEA RCT Registry (AEARCTR-0004630) and administered by FAIR-The Choice Lab.

1 Introduction

People sometimes make choices that are detrimental to their welfare. This creates opportunities for paternalistic interventions (Camerer, Issacharoff, Loewenstein, O'Donoghue, and Rabin, 2003; Thaler and Sunstein, 2003). The extent to which such opportunities should be used is a key issue in the relationship between the state and its citizens (Dworkin, 1972; Arneson, 1980; Le Grand and New, 2015). The role of paternalistic interventions is also at the heart of many interpersonal relationships, such as the relationship between parents and their children, experts and laypeople, employers and employees, and donors and recipients (Jacobsson, Johannesson, and Borgquist, 2007; Doepke and Zilibotti, 2017; Gangadharan, Grossman, Jones, and Leister, 2018; Kassirer, Levine, and Gaertig, 2020; Kiessling, Chowdhury, Schildberg-Hörisch, and Sutter, 2021; Beshears, Choi, Laibson, Madrian, and Skimmyhorn, 2022).

Opportunities for paternalistic interventions raise two fundamental normative questions. First, is it acceptable to restrict an individual's freedom to choose to promote their welfare? Second, who should judge whether an intervention improves the welfare of an individual? These questions have shaped an extensive normative literature, both in economics and in the social sciences more broadly, on how to justify paternalistic interventions (Nussbaum, 2001; Sen, 2004; Kaplow and Shavell, 2006; Hausman and McPherson, 2009; Le Grand and New, 2015; Thaler and Sunstein, 2021).

Libertarian Paternalism, which has become an influential position in recent years, can be characterized in terms of these two questions (Thaler and Sunstein, 2021). It only justifies interventions that (i) do not restrict people's freedom to choose and (ii) promote their welfare, as judged by themselves. Informed by insights from the behavioral sciences, a large and growing literature on libertarian paternalism has advocated interventions that manipulate the choice architecture without reducing the choice set of individuals, such as default options or information provision. These interventions aim to nudge people to make choices that are aligned with their own preferences (Thaler and Sunstein, 2021). The idea of libertarian paternalism 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).

The focus on the freedom to choose in libertarian paternalism contrasts the other main position in the normative literature, the classical *Welfarism* approach, which assesses policies only in terms of how they affect people's welfare (Sen, 2004). The welfarism approach finds interventions that restrict people's choice set acceptable as long as the interventions promote their welfare. The most prominent view of welfarism in economics overlaps with libertarian paternalism in considering an individual's own preferences to be the appropriate basis for welfare evaluations (Kaplow and Shavell, 2000, 2001). However, the welfarism approach can also be combined with other conceptions

of welfare that do not rely on satisfying an individual's own preferences (Nussbaum, 2001; Sen, 2004; Hausman and McPherson, 2009; Le Grand and New, 2015).

To shed light on people's paternalistic preferences and the extent to which they are consistent with libertarian paternalism or welfarism, we implement a novel experimental design in two large-scale experiments with more than 14,000 participants from the general population in the United States in the role of a spectator. In both experiments, the spectators make consequential decisions for another individual, the stakeholder. The first experiment, Study 1, explores whether a concern for the stakeholder's freedom to choose affects the spectator's willingness to intervene by manipulating the nature of the intervention. The second experiment, Study 2, investigates both the spectators' intervention decisions and their welfare evaluations, which allows us to provide estimates of the prevalence of libertarian paternalists and welfarists in the general population.

In Study 1, each spectator is matched with a stakeholder who will receive 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 a choice between the two options in a non-transparent choice environment. The spectator is informed that the non-transparent choice environment leads the stakeholder to be mistaken about the odds of the risky option and that, as a consequence, the stakeholder prefers the risky option to the safe option. The spectator is also informed that the matched stakeholder would prefer the safe option over the risky option if they were not mistaken about the odds of the risky option. In one set of treatments, the spectator is given the opportunity to implement a *hard intervention*, which removes the stakeholder's freedom to choose and gives them the safe option. In another set of treatments, the spectator is given the opportunity to implement a *soft intervention*, which does not restrict the stakeholder's choice set but informs the stakeholder about the correct odds of the risky option.

Study 1 also includes a second treatment dimension where we vary the reason why the stakeholder is mistaken about the choice set. A growing literature on social preferences documents 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 set of treatments, the spectator is informed that the stakeholder has made an incorrect calculation, which we refer to as a situation with *internal responsibility*. In another set of treatments, the spectator is informed that the stakeholder 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 intervene. 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, only about a third of the spectators are willing to implement the hard intervention that removes the stakeholder's freedom to choose. 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.

Taken together, Study 1 shows that, both for the internal and the external source of mistake, only a minority of the spectators are willing to implement the hard intervention, while a large majority are willing to implement the soft intervention. To guide the interpretation of this empirical pattern, we develop a theoretical framework in which a spectator's willingness to intervene is determined by two key features of paternalistic preferences: whether the spectator cares about the stakeholder's freedom to choose and how the spectator conceptualizes the stakeholder's welfare. We impose minimal assumptions on the spectator's behavior capturing that the spectators do not make dominated choices. The assumptions imply that the share of spectators whose welfare evaluations are aligned with the stakeholder's preferences should be (i) weakly larger than the share of spectators implementing the hard intervention and (ii) weakly smaller than the share of spectators implementing the soft intervention. Further, we formalize two paternalistic types, libertarian paternalists and welfarists, and show how the prevalence of these paternalistic types relates to key parameters in the theoretical framework.

The empirical pattern observed in Study 1 is consistent with the theoretical framework: the share of spectators implementing the hard intervention is smaller than the share of spectators implementing the soft intervention. However, Study 1 does not provide us with a measure of the share of spectators whose welfare evaluations are aligned with the stakeholder's preferences and thus does not allow us to fully test the two implications of the framework. Consequently, we conduct Study 2 that does not only manipulate the nature of the intervention but also elicits whether the spectator's welfare evaluation is aligned with the stakeholder's preferences. Specifically, Study 2 introduces a new treatment, in which the spectator decides whether to allocate the safe option or the risky option to the stakeholder, without having the opportunity to give the stakeholder the freedom to choose. Study 2 thus provides a stricter test of the theoretical framework and enables us to estimate the share of libertarian paternalists and welfarists based on both key features of paternalistic preferences.

Study 2 replicates the empirical pattern established in Study 1. We find again that a large majority of more than 80 percent of the spectators are willing to implement the soft intervention, while only about a third of the spectators are willing to implement the hard intervention. Hence, we establish in two independent large-scale samples of the general population in the United States that the nature of an intervention is of great

importance for the willingness to intervene. Further, in the new treatment, we find that a large majority of the spectators, about 70 percent, allocate the safe option to the stakeholder, aligned with the stakeholder's preferences. Taken together, the findings are in line with the theoretical framework: the share of spectators whose welfare evaluations are aligned with the stakeholder's preferences is (i) larger than the share of spectators implementing the hard intervention and (ii) smaller than the share of spectators implementing the soft intervention. Hence, Study 2 provides evidence that the willingness to intervene is determined by how an intervention affects the stakeholder's welfare and the stakeholder's freedom to choose. Finally, we estimate the prevalence of the two paternalistic types. We find that the majority of the spectators are welfarists, 52.9 percent, but also that a significant share of the spectators are libertarian paternalists, 34.4 percent. In terms of the two views on welfarism, we find that 35.4 percent of the spectators are welfarists whose welfare evaluations are aligned with the stakeholder's preferences and 17.5 percent are welfarists whose welfare evaluations are not aligned with the stakeholder's preferences. Overall, our findings show that the two paternalistic types can rationalize the behavior of most of the spectators in both Study 1 and Study 2.

Our results shed light on why soft interventions have gained strong support in recent years (Thaler and Sunstein, 2021). In line with the attention given to soft interventions in the policy debate (OECD, 2017), we find that the vast majority of the general population in the United States is willing to implement a soft intervention that promotes the welfare of a stakeholder, as judged by themselves. However, this should not be interpreted as evidence of most Americans being libertarian paternalists. In fact, in our studies, a large part of the support for the soft intervention comes from welfarists who are willing to implement both the soft intervention 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 set 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. As a result, even though we estimate libertarian paternalists to comprise only about a third of the population in the United States, 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 on the role of the nature of a paternalistic intervention for 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). Further, in a recent laboratory experiment with students, Ambuehl, Bernheim, and Ockenfels (2021) show that spectators project their own time preferences onto the stakeholders and are frequently willing to restrict the stakeholders' choice sets by removing impatient choice options. We extend this literature by showing in large-scale experiments with general population samples in the United States that the majority of people are willing to implement a soft intervention but not willing to implement a hard intervention. A novel feature of our setting is that the spectator is informed about the stakeholder's preferences, which allows us to observe the spectator's intervention decision in isolation of uncertainty about the stakeholder's preferences. Further, we provide evidence showing that the lower willingness to implement the hard intervention than the soft intervention is driven by a significant share of the population being libertarian paternalists who respect people's freedom to choose. Finally, we contribute to the literature by formalizing the two most prominent paternalistic types in the normative literature, libertarian paternalists and welfarists, and by providing estimates of the prevalence of these paternalistic types in a general population sample.

More broadly, the paper contributes to the literature examining heterogeneity in people's social preferences (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Konow, 2000; Andreoni and Miller, 2002; Charness and Rabin, 2002; Cappelen et al., 2007; Bellemare, Kröger, and van Soest, 2008; Cappelen, Konow, Sørensen, and Tungodden, 2013; Durante, Putterman, and van der Weele, 2014; Almås et al., 2020), by providing evidence of significant heterogeneity in people's paternalistic preferences. Our paper further contributes to the social preference literature by showing that the willingness to intervene does not depend on whether individuals are seen 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; Owens, Grossman, and Fackler, 2014; Sloof and von Siemens, 2017; Pikulina and Tergiman, 2020), by providing evidence that people do not only value their own autonomy but also respect other people's freedom to choose.

The paper is organized as follows. Section 2 presents Study 1. Section 3 presents the theoretical framework. Section 4 presents Study 2. Section 5 concludes. In the Appendix, we provide supplementary analysis, including adjustments for multiple hypothesis testing (Section A), the experimental procedures and instructions (Section B), and details about the pre-analysis plans (Section C).

2 Study 1

In this section, we present the experimental design, sample, empirical strategy, and main results of 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.

We first present the context of the intervention decision, before we introduce the treatment manipulations.

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 to the safe option. However, the stakeholder would prefer the safe option to 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 let 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 $\tilde{r} \succ s \succ r$.

The spectator is fully informed about the preference ranking of the stakeholder, and the experimental design builds on the assumption that the spectator believes that the stakeholder chooses according to their preferences. A preference-maximizing 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 $\tilde{r}$, and it holds that $\tilde{r} \succ s$. 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 that is not aligned with their preferences, and a spectator can intervene to ensure that they receive their preferred option.

Nature of Intervention. To study the role of the nature of the intervention for the spectators' intervention decision, 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. If the spectator implements the hard intervention, the stakeholder cannot make a choice but is allocated the safe option s. The soft intervention, in contrast, does not remove the stakeholder's freedom to choose. If the spectator implements the soft intervention, the stakeholder will make a choice in the transparent choice environment. The stakeholder will then know that the choice is between s and r. Taken together:

- The outcome of intervening, hard or soft, is that the stakeholder ends up with their preferred safe option s.
- The outcome of not intervening, hard or soft, is that the stakeholder ends up with their non-preferred risky option r.

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

Source of Mistake. To study the role of the source of mistake, the spectators are randomly assigned either to treatments in which the source of mistake is *internal* or to treatments in which the source of mistake is *external*. In treatments with the internal source of mistake, 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 with the external source of mistake, the spectator is informed that they are matched to a stakeholder who was unlucky and received incorrect information about the odds of the risky option. The source of mistake does not affect the choices of a preference maximizing stakeholder and, consequently, it does not affect the outcomes of intervening or not intervening.

It follows that if the spectators' willingness to intervene does not depend on the source of mistake, the share of spectators that intervene when the source of mistake is internal would be equal to the share of spectators that intervene when the source of mistake is external. Hence, the experimental design allows us to identify whether the source of mistake matters for the spectators' willingness to intervene, and whether spectators are more willing to intervene when the source of mistake is internal or external.

Treatment Design. We implemented a full factorial 2×2 between-subjects design. We refer to the four treatments as *Hard*×*Internal*, *Hard*×*External*, *Soft*×*Internal*, and

Soft×*External*. Figure 1 summarizes the sequence of the main events in the experiment and shows at which stage the treatment manipulations come into play.

[Figure 1 about here]

2.2 Participants

In this section, we present the sample of spectators and explain how we recruited and matched stakeholders to spectators.

Spectators. The spectators were recruited from the general population in the United States through a professional data service company (Dynata). We sampled a total of 8,004 spectators in August 2019, based on quotas for gender, age, education, income, and region, to match a representative sample of the general population in the United States (aged 18 or older). The 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.

We elicited the spectators' demographic background characteristics, including gender, age, education, and income. Further, since the intervention decision is made in a context where the stakeholder is about to make a choice in the domain of risk, we also measured the spectators' own willingness to take risks by eliciting their self-assessment on an 11-point scale ranging from "Completely unwilling to take risks" (0) to "Very willing to take risks" (10). At the end of the experiment, spectators could self-identify as "Republican," "Democrat," or "Independent/Third Party." Finally, we elicited the extent to which the spectators agree 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." The spectators could answer on a seven-point scale ranging from "fully disagree" (1) to "fully agree" (7).

Table 1 reports the main demographic characteristics of the sample in Study 1 and compares it to the population in the United States. From the leftmost column, we observe that the sample is gender-balanced and exhibits significant heterogeneity in age, education, and income. The median age in our sample is 46 years. In terms of education, 31% of the sample do not have any college education, while 15% have a Master's degree or an even higher educational attainment. The median educational category is "some college." About one quarter of the sample has a yearly income of less than USD 30,000, while slightly more than 10% have an income exceeding USD 150,000. The median income category is USD 30,000 to 60,000. 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. We also observe substantial heterogeneity in the spectators' risk preferences: 23% of the spectators indicate that their willingness to take risks is below the midpoint of 5 on the 11-point scale from 0 to 10, and 64% indicate that their willingness to take risks is above the midpoint; the median response is 6. Comparison of the sample with the population averages in the United States (rightmost column) reveals that the 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. Table A1 in the Online Appendix shows that the sample is balanced across treatments.

[Table 1 about here]

Stakeholders. We recruited the stakeholders through an online labor market platform (Amazon Mechanical Turk). The stakeholders could receive a bonus payment. We elicited the stakeholders' preferences over the safe and the risky bonus option in both the transparent choice environment and in the non-transparent choice environment. 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 a spectator.[1]

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:

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

The dependent variable I_i is an indicator for whether spectator i intervenes. Treatment *Hard*×*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_iE_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

[1] See Section B.2 of the Appendix for further details about the preference elicitation and the matching protocol.

the sample. Female indicates whether a spectator is female or male. We estimate the models with and without the vector of background characteristics.

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

We start by providing an overview of the spectators' intervention decisions. Figure 2 shows the share of spectators that intervene by treatment. The left panel shows the share of spectators that intervene in treatments *Hard*×*Internal* and *Soft*×*Internal*. The right panel shows the share of spectators that intervene in treatments *Hard*×*External* and *Soft*×*External*.

[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, 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 is allocated their preferred safe option. In contrast, a large majority of the spectators implement the soft intervention (internal: 85.9 percent, external: 87.5 percent), which preserves the stakeholder's freedom to choose and ensures that the stakeholder can make a choice in the transparent choice environment.

[Table 2 about here]

Table 2 reports the regression analysis. In column (1), pooling the treatments with the hard intervention and the treatments with the soft intervention, 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 causal effect is virtually unaffected by 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.

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 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, $p < 0.01$).

We summarize the analysis of how the nature of the intervention affects the spectators' willingness to intervene as follows:

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 mistake on the spectators' willingness to intervene. Columns (3) and (4) show that the estimated share of spectators that implement the hard intervention is 3.5 percentage points lower when the source of mistake is external rather than internal ($p < 0.05$), while the estimated share of spectators that implement the soft intervention is 1.6 percentage points higher, but the difference is not statistically significant ($p = 0.15$ and $p = 0.13$, respectively). In columns (5) and (6), pooling the treatments with the internal source of mistake and the treatments with the external source of mistake, we find that there is no significant

average effect of the source of mistake on the spectators' willingness to intervene. Taken together, we conclude:

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.*

The fact that we have a large-scale general population sample allows us to study whether different subgroups of the population differ in their intervention decisions. In Figure 3, we report the heterogeneity analysis for subgroups defined by political orientation, willingness to take risks, education, income, age, and gender. The left panel shows the estimated interaction effect between an indicator for the spectator being in one of the treatments with the soft intervention and an indicator for the respective background characteristic. The right panel shows the estimated interaction effect between an indicator for the spectator being in one of the treatments with external source of mistake and an indicator for the respective background characteristic.

[Figure 3 about here]

In the left panel of Figure 3, we observe that the estimated average causal effect of the nature of the intervention on the spectators' willingness to intervene, both for the single and the joint interaction model, is not significantly different across subgroups, with the exception of High Education and High Income. The difference between above-median and below-median educated spectators is significant in both the single interaction model and the joint interaction model and robust to multiple hypothesis testing (see Tables A4 and A7 in the Appendix). In the single interaction model, the estimated average causal effect is 6.0 percentage points larger for above-median educated spectators than for below-median educated spectators ($p < 0.01$). Above-median educated spectators are slightly less likely to implement the hard intervention (30.4 vs. 32.9 percent) and slightly more likely to implement the soft intervention (88.5 vs. 85.0 percent) than below-median educated spectators. The difference between spectators with above-median and below-median income is significant only in the single interaction model but not in the joint interaction model, and it is not robust to multiple hypothesis testing. The right panel shows that the estimated average causal effect of the source of mistake on the spectators' willingness to intervene is not different across subgroups (see Tables A5 and A8 in the Appendix).

There is no significant interaction effect with respect to political orientation, both in terms of the nature of the intervention and the source of mistake. Further, taking into account that we do not find a significant level effect of being Republican (see Table 2),

it follows that Republicans and non-Republicans make very similar intervention decisions in the experiment.[2] This finding may 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 to fundamental differences in paternalistic preferences.

The heterogeneity analysis also shows that the estimated average causal effect of the nature of the intervention is large and highly significant in all subgroups ($p < 0.01$ in all tests, see Table A4). In contrast, the estimated average causal effect of the source of mistake is small and not significant in all subgroups ($p > 0.14$ in all tests, see Table A5).

We sum up the heterogeneity analysis in the following result:

Result 3: *There are only small differences in 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

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

We introduce a theoretical framework to shed light on these questions.

[2]Table A6 shows that this finding is robust to focusing on the sub-sample of Republicans and Democrats and to different model specifications.

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, $c(ni) = c^{+nt}$. Hence, 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 intervenes if and only if $F(c(i)) \geq F(c(ni))$ and $W(b(i)) > W(b(ni))$, with $W(b) = U(b)$.

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 explained 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

We first describe the experimental design, sample, and empirical strategy, before we present and discuss the main results of Study 2.

4.1 Experimental Design and Participants

Experimental Design. Study 2 replicates treatments *Hard*×*Internal* and *Soft*×*Internal* from Study 1, which, for short, we refer to as treatments *Hard* and *Soft* in the following. Study 2 adds a treatment, labeled *Welfare*. The context of treatment *Welfare* is identical to the context of treatments *Hard* and *Soft*. In contrast to treatments *Hard* and *Soft*, a spectator in treatment *Welfare* does not have the option to give the stakeholder the freedom to choose, but must allocate either the preferred safe option or the non-preferred risky option to the stakeholder. We use treatment *Welfare* to directly elicit whether a spectator considers the safe option or the risky option to promote the welfare of the stakeholder.

Participants. The 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 a total of 6,033 spectators in January 2020. The middle column of Table 1 reports the characteristics of the sample in Study 2, which is very similar to the sample in Study 1. Table A9 shows that the sample is largely balanced across treatments, with slightly fewer Republicans and slightly more spectators with a higher education and a higher income in treatment *Welfare*.

4.2 Empirical Strategy

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

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

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

We start by providing an overview of the spectators' decisions. Figure 4 shows the share of spectators that intervene in treatments *Hard* and *Soft*, respectively, and the share that allocates the preferred safe option to the stakeholder in treatment *Welfare*. We observe from treatments *Hard* and *Soft* that the empirical pattern from Study 1 replicates in Study 2. 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 4 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 4 about here]

Table 3 reports the regression analysis. In column (1), we observe that 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* ($p < 0.01$). We further estimate that the share of spectators that intervene in treatment *Hard* is 34.4 percentage points lower than the share of spectators that allocate the safe option in treatment *Welfare* ($p < 0.01$). Column (2) shows that the estimated treatment differences are virtually unaffected by the inclusion of the spectators' background characteristics.

[Table 3 about here]

The estimates in Table 3 are in line with Observation 1: (i) 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 ($\hat{\theta}^H < \hat{\theta}^A$), and (ii) 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 ($\hat{\theta}^A < \hat{\theta}^S$).

Figure 5 shows that Observation 1 holds across subgroups of the general population. In each subgroup, we observe that the estimated share of spectators that implement the hard intervention is strictly smaller 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 (see Table A12).[3]

[Figure 5 about here]

We can now state the following result:

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.*

Given that the spectators' intervention decision are consistent with Observation 1, we can use the estimates in Table 3 to study the prevalence of the two main paternalistic types in our sample. Based on column (1), 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

[3]Figure 5 also shows that the share of spectators allocating the preferred safe option to the stakeholder is largest among below-median risk takers and smallest among above-median risk-takers. This suggests that some stakeholders rely on their own preferences in their welfare evaluations (Ambuehl et al., 2021).

whose welfare evaluations are not aligned with the stakeholder's preferences, $\hat{\sigma}^{W_{na}}$, is 17.5 percent. We summarize the analysis of the prevalence of the main paternalistic types as follows:

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.*

The estimation results imply that only the decisions of 12.7 percent of the spectators (given by $\hat{\theta}^S - \hat{\theta}^A$) cannot be explained by the two paternalistic types. 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. Overall, given that both these spectators and the libertarian paternalists value the stakeholder's freedom to choose, we estimate that for almost half of the sample, 47.1 percent, the willingness to intervene is determined not only by how an intervention affects the stakeholder's welfare but also by how an intervention affects the stakeholder's freedom to choose.

Finally, Figure 5 shows that the prevalence of the two paternalistic types is quite similar across subgroups. In each subgroup, we find that about half of the spectators are welfarists, but also that a significant share of spectators are libertarian paternalists.

5 Conclusions

The paper studies paternalistic preferences in two large-scale experiments with participants from the general population in the United States. The experimental context 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 own preferences, and a spectator can intervene to ensure that they receive their preferred option. Our experimental study provides, in two independent samples, causal evidence for the nature of an intervention being of great importance for the willingness to intervene. 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 the choice set – whether the source of mistake is internal or external – and in different 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 the two paternalistic preference types: about a third of the spec-

tators are estimated to be libertarian paternalists and about half of the spectators are estimated to be welfarists. The estimated share of libertarian paternalists reveals that a significant part of the population in the United States is reluctant to implementing hard interventions: they prefer leaving people "free to fail," rather than enhancing their welfare by restricting their choice options.

To study whether people consider that this experimental context is relevant for the policy debate on paternalism in the United States, we also asked two general questions. First, we asked the spectators whether they agree that people sometimes make choices that are harmful to themselves. We find that the large majority of the spectators agree with the statement (see upper panel of Figure A1). Second, we asked the spectators 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 that almost half of the spectators express some 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 (e.g., Kuziemko, Norton, Saez, and Stantcheva, 2015) but also that people consider the freedom to choose to be an integral component of a good life. Interestingly, we find that the share of people disagreeing that the government can improve people's lives with hard paternalism is very close to the estimated share of spectators that value the stakeholder's freedom to choose in our experiment, 47.0 percent vs. 47.1 percent. Further, we find a strong positive association between abstaining from implementing the hard intervention in our experiment and disagreement with the view that the government can improve people's lives by means of hard paternalism ($p < 0.001$). Taken together, we find evidence that the spectators perceive the experimental context of the present study to be of relevance and that the willingness to intervene in the experiment is predictive of people's attitudes toward governmental paternalistic policies.

The experimental paradigm developed in this paper allows addressing a multitude of intriguing questions on paternalistic preferences. First, it is important to study the extent to which paternalistic preferences are domain-specific. Are paternalistic preferences different when stakeholders make choices over time than when they make choices in the domain of risk? Second, what is the role of the relation between the spectator and the stakeholder? A growing literature has focused on the economics of parenting (Francesconi and Heckman, 2016; Doepke and Zilibotti, 2017; Cobb-Clark, Salamanca, and Zhu, 2019; Agostinelli, Doepke, Sorrenti, and Zilibotti, 2020), which raises the question of whether people make different paternalistic considerations when acting as parents. Finally, it would be interesting to study cultural variation in paternalistic preferences, and the extent to which this variation can contribute to explain cross-country differences in people's attitudes to paternalistic policies (Sunstein et al., 2018). Paternalistic interventions are prevalent across societies, and it is of great importance to understand how they are justified and relate to people's own paternalistic preferences.

References

Agostinelli, Francesco, Matthias Doepke, Giuseppe Sorrenti, and Fabrizio Zilibotti (2020). "It takes a village: The economics of parenting with neighborhood and peer effects," Working Paper 27050, National Bureau of Economic Research.

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.

Beshears, John, James J. Choi, David Laibson, Brigitte C. Madrian, and William L. Skimmyhorn (2022). "Borrowing to save? The impact of automatic enrollment on debt," *The Journal of Finance*, 77(1): 403–447.

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.

Cappelen, Alexander W., James Konow, Erik Ø. Sørensen, and Bertil Tungodden (2013). "Just luck: An experimental study of risk taking and fairness," *American Economic Review*, 103(3): 1398–1413.

Charness, Gary and Matthew Rabin (2002). "Understanding social preferences with simple tests," *Quarterly Journal of Economics*, 117(3): 817–869.

Cobb-Clark, Deborah A., Nicolas Salamanca, and Anna Zhu (2019). "Parenting style as an investment in human development," *Journal of Population Economics*, 32(4): 1315–1352.

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

DellaVigna, Stefano and Elizabeth Linos (2022). "RCTs to scale: Comprehensive evidence from two nudge units," *Econometrica*, 90(1): 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).

Doepke, Matthias and Fabrizio Zilibotti (2017). "Parenting with style: Altruism and paternalism in intergenerational preference transmission," *Econometrica*, 85(5): 1331–1371.

Durante, Ruben, Louis Putterman, and Joël J. van der 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.

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(1): 1–10.

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.

Francesconi, Marco and James J. Heckman (2016). "Child Development and Parental Investment: Introduction," *The Economic Journal*, 126(596): F1–F27.

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): 62–74.

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.

Kassirer, Samantha, Emma E Levine, and Celia Gaertig (2020). "Decisional autonomy undermines advisees' judgments of experts in medicine and in life," *Proceedings of the National Academy of Sciences*, 117(21): 11368–11378.

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.

Nussbaum, Martha (2001). "Symposium on amartya sen's philosophy: Adaptive preferences and women's options," *Economics and Philosophy*, 17(1): 67–88.

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

Owens, David, Zachary Grossman, and Ryan Fackler (2014). "The control premium: A preference for payoff autonomy," *American Economic Journal: Microeconomics*, 6(4): 138–161.

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 (2004). *Rationality and Freedom*, Harvard University Press.

Sloof, Randolph and Ferdinand A. von Siemens (2017). "Illusion of control and the pursuit of authority," *Experimental Economics*, 20: 556–573.

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.

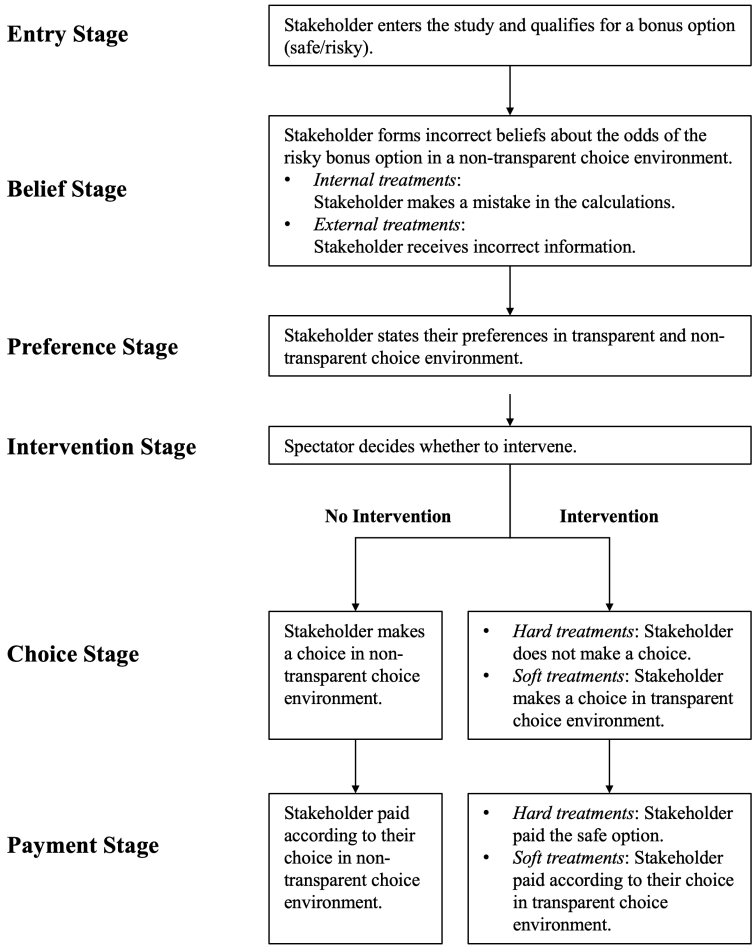
Tannenbaum, David, Craig R. Fox, and Todd Rogers (2017). "On the misplaced politics of behavioral policy interventions," *Nature Human Behaviour*, 1: 0130.

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

Thaler, Richard H. and Cass R. Sunstein (2021). *Nudge: The Final Edition: Improving Decisions About Money, Health, and the Environment*, New Haven: Yale University Press.

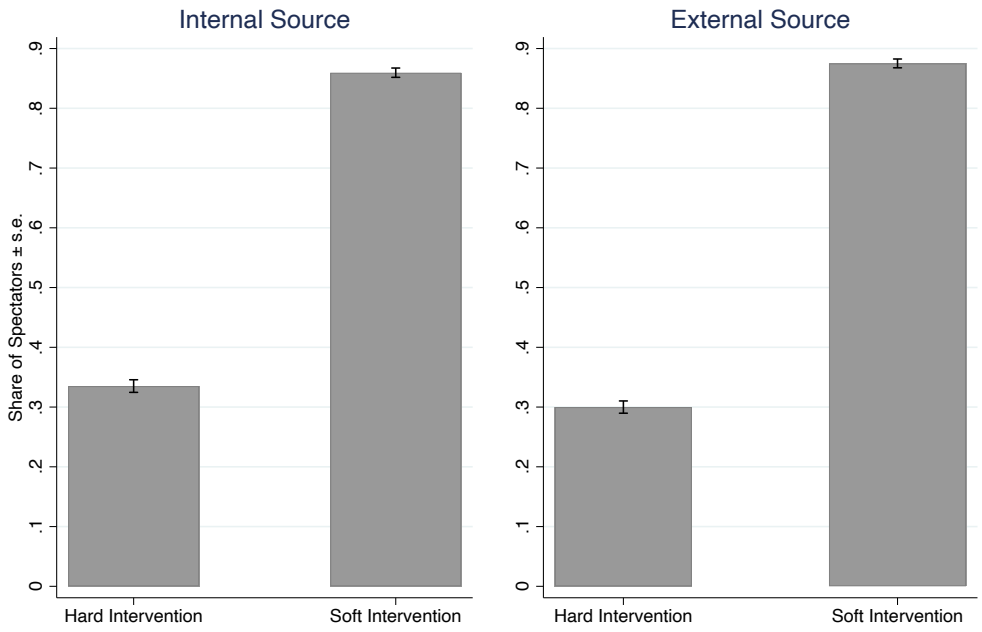
US Census Bureau (2018). "Data from www.census.gov, 2018–2020," Library Catalog: www.census.gov Section: Government.

Figure 1: Sequence of Events in Study 1



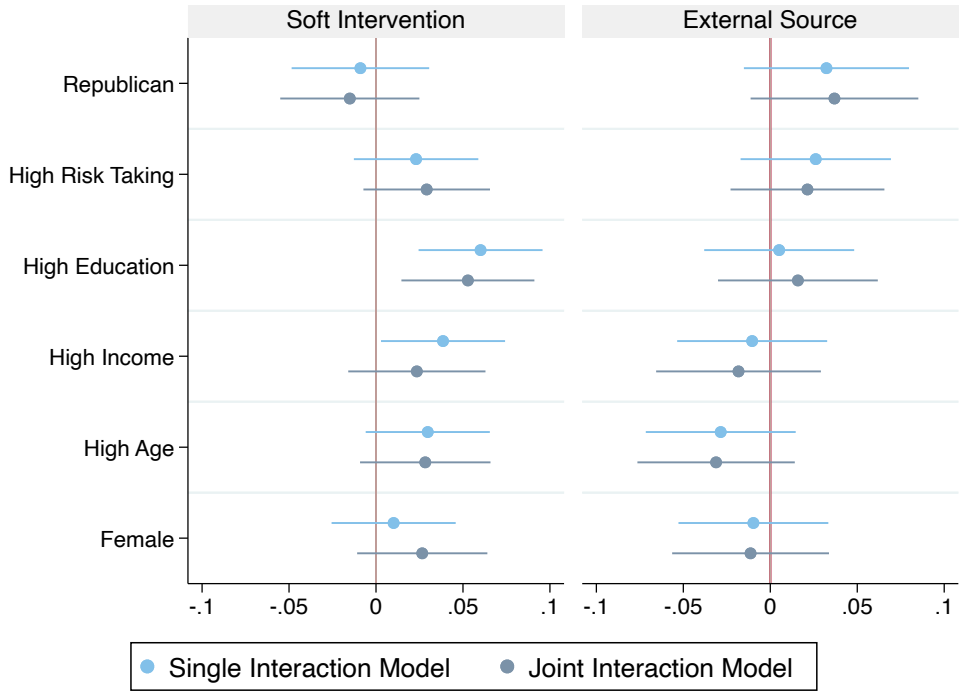
Notes: The figure shows the events that take place in each stage of Study 1.

Figure 2: Spectator Decisions by Treatment — Study 1



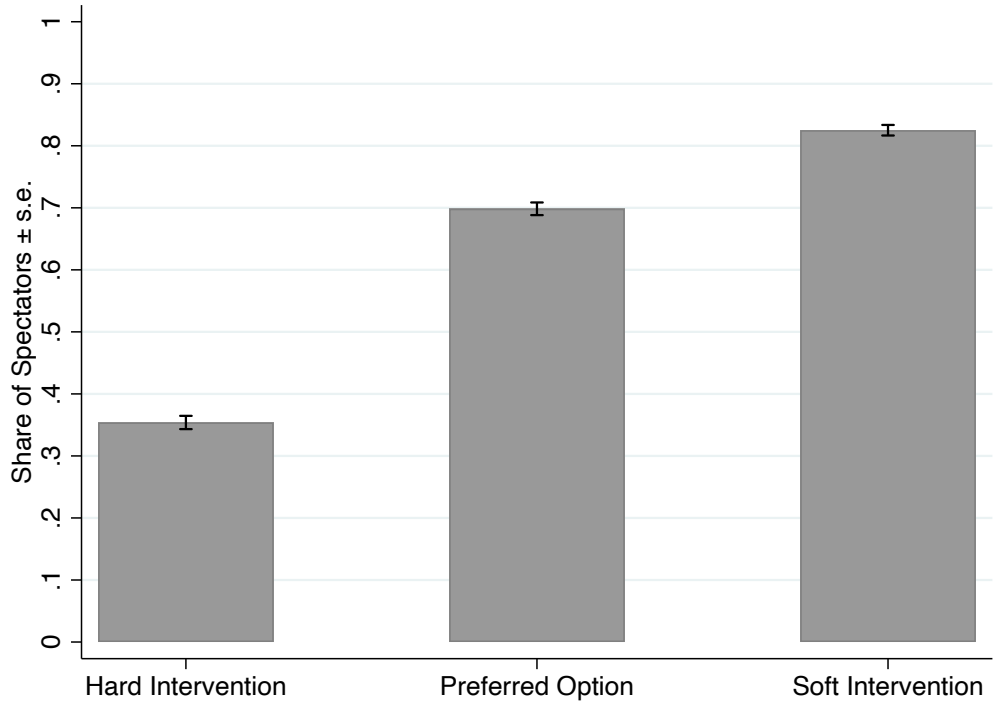
Note: The figure shows the share of spectators that intervene by treatment. The left panel shows the share of spectators intervening in treatments *Hard*×*Internal* and *Soft*×*Internal*. The right panel shows the share of spectators intervening in treatments *Hard*×*External* and *Soft*×*External*.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

Figure 3: Heterogeneous Treatment Effects — Study 1



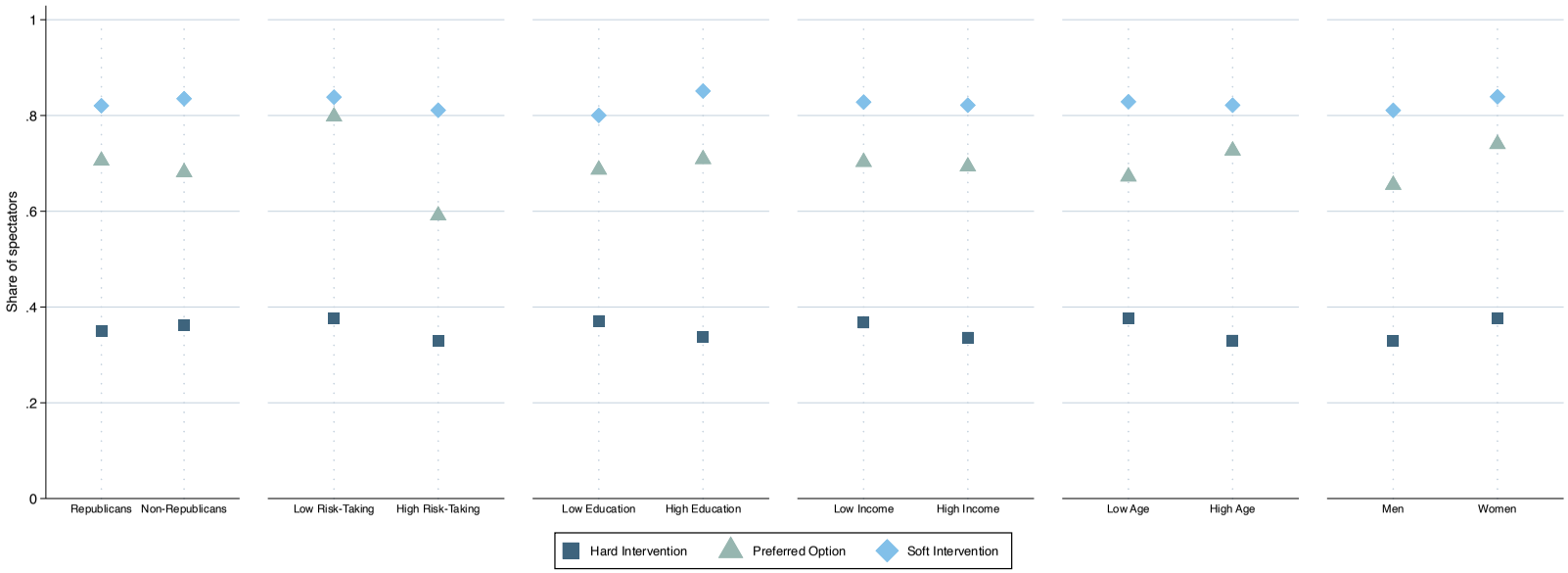
Note: The figure shows the estimated coefficients and 95% confidence intervals for the interaction effects, both in the single interaction models and the joint interaction model. The left panel shows the estimated interaction effect between an indicator for the spectator being in treatment *Soft*×*Internal* or *Soft*×*External* and an indicator for the respective background characteristic. The right panel shows the estimated interaction effect between an indicator for the spectator being in treatment *Soft*×*External* or *Hard*×*External* and an indicator for the respective background characteristic. See Tables A4 and A5 in Appendix A for the underlying regression results.

Figure 4: 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 5: Spectator Decisions by Subgroup — Study 2



Note: The figure shows, 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*. See Table A12 for the underlying regression results.

Table 1: 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 2: 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 Tables A2 and A3 in Appendix A). Robust standard errors in parentheses. * $p<.10$, ** $p<.05$, *** $p<.01$

Table 3: Regression Results — Study 2

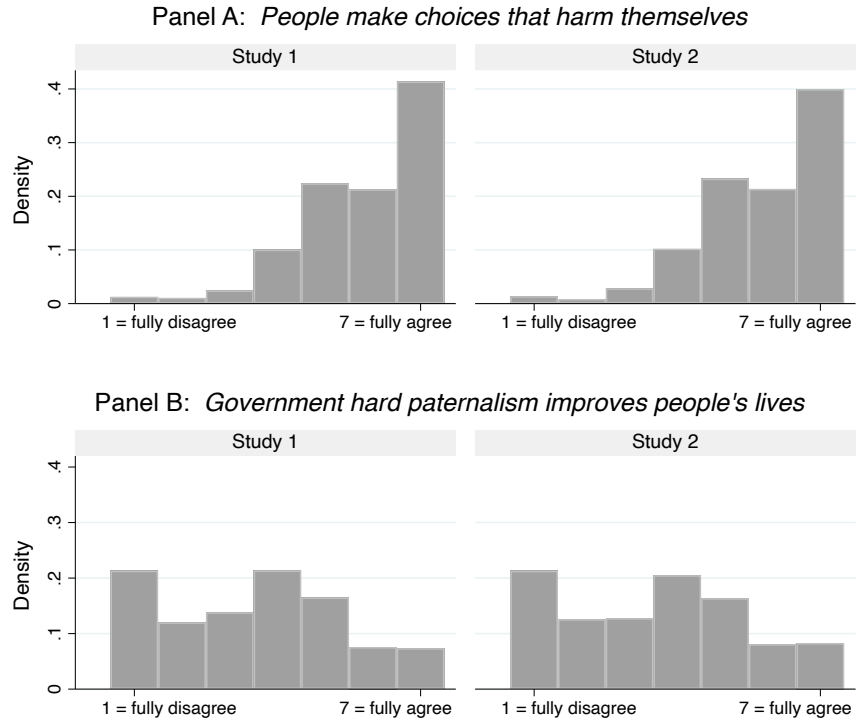
	(1)	(2)
Soft Intervention	0.127***	0.128***
	(0.013)	(0.013)
Hard Intervention	-0.344***	-0.344***
	(0.015)	(0.015)
Republican		0.012
		(0.012)
High Risk Taking		-0.094***
		(0.012)
High Education		0.030**
		(0.012)
High Income		-0.009
		(0.013)
High Age		-0.015
		(0.012)
Female		0.050***
		(0.012)
Constant	0.698***	0.711***
	(0.010)	(0.017)
Observations	6033	6033
R^2	0.169	0.182

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 in models (1) and (2) that include the data from all three treatments. Model (3) comprises only the data from treatment *Welfare*. "Soft Intervention" and "Hard Intervention" are indicators 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. The results are robust to adjusting for multiple-hypothesis testing and to using Probit models (see Tables A10 and A11 in Appendix A). Robust standard errors in parentheses. * $p<.10$, ** $p<.05$, *** $p<.01$

Online Appendix

A Additional Figures and Tables

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 ($n = 6,014$).

Table A1: 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 A2: MHT Corrections for Table 2

	(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 A3: Probit Models for Table 2

	(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 2. Robust standard errors in parentheses. * $p<.10$, ** $p<.05$, *** $p<.01$

Table A4: 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 A5: 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 A8). Robust standard errors in parentheses. * $p<.10$, ** $p<.05$, *** $p<.01$

Table A6: 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 Democrates. 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. * $p<.10$, ** $p<.05$, *** $p<.01$

Table A7: MHT Corrections for Table A4

	(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: MHT Corrections for Table A5

	(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 A9: 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. * $p<.10$, ** $p<.05$, *** $p<.01$

Table A10: MHT Corrections for Table 3

	(1)	(2)
Soft Intervention	0.127***	0.128***
	[.001]	[.001]
Hard Intervention	-0.344***	-0.344***
	[.001]	[.001]
Republican		0.012
High Risk Taking		-0.094***
High Education		0.030**
High Income		-0.009
High Age		-0.015
Female		0.050***
Constant	0.698***	0.711***
Observations	6033	6033
R^2	0.169	0.182

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 in models (1) and (2). Model (3) comprises only the data from treatment *Welfare*. "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. 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 Column (1) and within Column (2), respectively. Adjusted p-values are reported in brackets. * $p<.10$, ** $p<.05$, *** $p<.01$

Table A11: Probit Models for Table 3

	(1)	(2)
Soft Intervention	0.415***	0.421***
	(0.044)	(0.044)
Hard Intervention	-0.895***	-0.908***
	(0.041)	(0.041)
Republican		0.037
		(0.039)
High Risk Taking		-0.289***
		(0.036)
High Education		0.099**
		(0.038)
High Income		-0.030
		(0.039)
High Age		-0.046
		(0.037)
Female		0.156***
		(0.037)
Constant	0.520***	0.563***
	(0.029)	(0.052)
Observations	6033	6033

Notes: The table reports probit models corresponding to the OLS models shown in Table 3. Robust standard errors in parentheses. * $p<.10$, ** $p<.05$, *** $p<.01$

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

B.1 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.1.1 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.1.2 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.1.3 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.1.4 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.1.5 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.

B.2 Stakeholders

Here we provide further 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:

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, some stakeholders received a signal that would allow a Bayesian individual to correctly calculate that the likelihood of receiving USD 10 is 50% (internal condition). 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*

Other stakeholders received an incorrect signal about the likelihood of receiving USD 10 and were informed that the average of all signals sent is correct (external condition). 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, 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. 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.

Given the 5:1 matching between spectators and stakeholders, we recruited stakeholders until we reached the necessary number of 1,601 stakeholders who could be matched with a spectator for Study 1 and 1,207 stakeholders who could be matched with a spectator for Study 2.

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 Tables 2 and 3 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 2 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 2. 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 3 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 3. 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. * $p<.10$, ** $p<.05$, *** $p<.01$