Tontrup, Stephan; Arlen, Jennifer; Sprigman, Christopher Jon

Working Paper

Behavioral Self-Management and the Strategic Shifting of Fairness Norms

Behavioral Self-Management and the Strategic Shifting of Fairness Norms

Stephan Tontrup, Jennifer Arlen, & Christopher Sprigman[1*]

Abstract

People often act prosocially and voluntarily conform to social and legal norms. This has fueled the idea that law can guide behavior through its expressive power. By contrast, we offer a theoretical and experimental framework suggesting that people strategically alter their decision-making environment to shift the norm applicable to their actions to one that is in their self-interest and to the detriment of others. Norm-shifting is one strategy within a broader concept we refer to as Behavioral Self-Management (BSM).

To test norm-shifting, we implement a dictator game in which Allocators are offered an effort task before allocating a sum between themselves and a Recipient. Allocators receive the same endowment whether or not they work. We hypothesize that many will undertake the task to shift the applicable fairness norm from equal division to an effort-based norm that justifies their retaining a larger share. Prior evidence shows that costly effort is widely perceived as legitimizing unequal outcomes.

We find that many Allocators decide to work, thereby reducing average transfers. Their work choices are strategic: their odds of working are higher the more they expect work to shift the fairness norm in their favor and the more prosocial they are—that is, the higher the moral costs they face for violating the fairness norm. Finally, Allocators who work make transfers that they expect to conform to an effort-based norm in the view of others, to maintain their self- and social-image.

Our findings have implications for compliance with the law and with social norms. BSM can enable selfish non-compliance by undermining the social norms that underpin the law or by establishing social norms that provide justification for violation, while avoiding the social disapproval that would otherwise result.

[1*] Stephan Tontrup is the Lawrence Jacobson Fellow of Law and Business; Jennifer Arlen is the Norma Z. Paige Professor of Law, Faculty Director of the Program on Corporate Compliance & Enforcement and the Center on Law, Economics & Organization, and Christopher Sprigman is the Murray and Kathleen Bring Professor of Law, and Co-Director, Engelberg Center on Innovation Law and Policy, all at the New York University School of Law. We would like to thank New York University School of Law and the Filomen D'Agostino and Max E. Greenberg Research Fund for their financial support for this project and the following for their helpful comments: Gilat Bachar, Oren Bar-Gil, Luca Baltensperger, Stefan Bechtold, Richard Brooks, Christoph Engel, Yuval Feldman, Wolfgang Gaissmaier, Talia Gillis, Elysa Dishman, Samuel Estreicher, Franco Ferrari, Claire Hill, Daniel Hemel, Yoan Hermstrüver, Geeyoung Min, Hajin Kim, Tamar Kricheli-Katz, Mattias Kumm, Moran Ofir, Benjamin Scheibehenne, Brian Sheppard, Holger Spamann, Roseanne Summers, Andrew Verstein, Tom Zur, Eyal Zamir, and participants at the 2025 annual meeting of the American Law & Economics Association; the 2025 Conference on Empirical Legal Studies; the 2025 Asian Law and Economics Association annual meeting; the University of Chicago Law & University of Michigan Law PALS workshop; the 2nd NYU Workshop at Kyoto University Graduate School of Law; the 12th International Meeting in Law & Economics in Paris; the Workshop Series at the ETH Zürich, the Center on Law and Social Science Faculty Workshop, the Internal Law and Economics Faculty Workshop, NYU School of Law; the University of Southern California Gould School of Law; the Experimental Economics Workshop at the Economics Department of George Mason University; the VCOMP webinar on Innovations in Compliance Research, the Bar Ilan University; and the Virtual Compliance Workshop, Duke Law School.

I. Introduction

To achieve their goals, policymakers seeking to shape behavior via changes in legal rules must base their interventions on a reasonably accurate theory of how people are likely to respond. Behavioral legal scholars have sought to supply such a framework. Rather than assume that people invariably act in their individualistic self-interest, these scholars rely on empirical evidence that people have other-regarding preferences which lead them to comply with fairness, social, or legal norms even when doing so is contrary to their self-regarding self-interest (Ostrom, 2000; Blader & Tyler, 2003; Fehr & Fischbacher, 2004; Bicchieri, 2006; Tyler, 2021; Bowles & Gintis, 2011). People with stronger other-regarding preferences have also been shown to care more about their social-image and therefore tend to comply more with fairness and social norms (Ariely et al., 2009; Regner, 2021; Tontrup & Sprigman, 2025). For some, other-regardingness is also an internalized part

of their self-image that motivates them to comply with norms (Andreoni & Bernheim, 2009; Falk, 2021; Gross & Vostroknutov, 2022; Matthey & Regner, 2014; Regner, 2021).

Relying on this evidence that people sacrifice self-interest for other-regarding concerns, behavioral legal scholars have offered a host of policy proposals. For example, some argue that people's inclination to comply with social norms, and also legal injunctions, is sufficiently strong that the law can deter by relying on its ability to express social condemnation of violations and need not always resort to formal criminal punishment, which is socially costly (Bilz & Nadler, 2014; Dau-Schmidt, 1990; Cooter, 2000; Sunstein, 1986; Sunstein, 1996a; Jolls & Sunstein, 1998; Tyler, 2021; McAdams, 1997). For example, some have suggested imposing duties on corporate directors unaccompanied by serious risk of sanction for breach on the grounds that corporate executives are motivated to comply with norms, such as fiduciary duties (Stout 2002, 2011; Dorff, 2003; see also Cooter & Eisenberg, 2001; Greenfield, 2001; for a general critique see Bubb & Pildes, 2014).

What is Behavioral Self-Management (BSM). By contrast, our work suggests that policymakers should remain cautious and cannot necessarily rely on social and legal norms to constrain self-interest in the absence of effective enforcement. In this Article, we provide theory and experimental evidence that people who are otherwise prosocial will restructure their decision-making environment in order to act in their self-interest without damage to their self- and social-image as a norm-complier—an example of what we have labeled Behavioral Self-Management (BSM). For example, decision-makers may delegate an action to an agent when the action is privately beneficial but also risks violating social or ethical norms. By having the agent carry out the action, they reduce the degree of responsibility attributed to themselves for the violation and lower the self- and social-image costs they incur (Hamman, Loewenstein & Weber, 2010; Tontrup & Sprigman, 2025; see also Arlen & Tontrup, 2015). Another BSM strategy we currently study is the access of information about the compliance of others when considering a self-interested norm breach. Learning that others fail to comply can reduce self-image costs, since one's own behavior does not appear to be worse by comparison, and it can diminish social-image costs, as others who themselves do not comply with the norm are less likely to condemn a violation.

One central implication of BSM theory is that we expect many people, in particular those who are otherwise prosocial, to behave self-interestedly in real-world settings that provide opportunities to engage in BSM strategies. In such settings, people can ease the constraints that fairness concerns, and that legal, and social norms, place on them. By identifying and using BSM strategies, individuals change the rules of the game in their favor, as we will show in this article. This contrasts with standard economic games often used to analyze the effect of prosocial preferences on behavior; these settings do not provide such BSM opportunities.

Our premise that people tend to actively evade the constraints that norms place on their pursuit of individualistic self-interest is supported by a substantial body of evidence on passive forms of evasion. For example, people exhibit motivated reasoning, acquiring and processing information in ways that support desired self-interested choices (Kunda, 1990; see also Bazerman & Tenbrunsel, 2006; see generally Arlen & Kornhauser, 2023). They also display self-serving bias, interpreting fairness so as to favor their own outcomes (Babcock, Loewenstein, Issacharoff & Camerer, 1995). Individuals exploit "moral wiggle room" to justify self-interested choices by deliberately remaining

ignorant of information that would make the harm their choices impose on others salient (Dana et al., 2007; Bazerman & Tenbrunsel, 2011; Grossman 2014; Fahrenwaldt et al., 2024; Offer et al., 2024).

Cornerstones of BSM. Our BSM theory builds on this literature, but differs from it in three core elements: We focus on the *active, strategic steps* decision-makers take to reshape their objective decision environment—through delegation, through seeking information on others' compliance, or, as in this study, through norm-shifting—in order to mute the self- and social-image costs of favoring themselves.

First, BSM focuses on strategies that influence not only the actor's self-perception but also the normative judgments of others. This contrasts with mechanisms such as moral wiggle room or motivated reasoning, where decision-makers primarily manipulate their own beliefs about the normative value of their behavior. By shifting how others judge the behavior, BSM strategies allow decision-makers to lower their social-image costs in addition to their self-image costs. In the moral-wiggle-room literature, by contrast, decision-makers can at best preserve their self-image—and may do so at the expense of their social-image. This makes BSM a more attractive and potentially more effective strategy.

Second, BSM is strategic behavior. In a game-theoretic sense, strategic behavior refers to actions in interdependent environments where a person's best choice depends on others' beliefs and responses. Strategic actors choose actions to influence these beliefs and responses toward their preferred outcomes. Behaviorally rational actors optimize given their social preferences or other behavioral motives (Camerer, 2003; Fehr & Schmidt, 1999; Battigalli & Dufwenberg, 2022). BSM is strategic in precisely this sense. When individuals engage in strategic norm-shifting, they anticipate that if others accept a less demanding fairness norm, they will not condemn their self-serving behavior. Actors therefore shift norms to influence other peoples' normative judgments and avoid social-image costs. The same strategic logic operates internally: individuals anticipate their future self's moral self-assessment and shift norms to influence this judgment, avoiding the self-image costs that would arise from violating a fairness norm.

Others may also behave strategically toward the norm shifter. They may impose condemnation to prevent norm-shifting when rejecting the new norm better preserves their own moral self-image and social standing. Alternatively, they may accept or even reward norm-shifting to exploit the lower normative standard themselves, pursuing self-interest at reduced image costs.

This focus on strategic behavior distinguishes BSM from existing literature and matters for its practical relevance. By shifting which norm applies or what it prescribes, actors can establish a new normative status quo that benefits them and create opportunities that would not otherwise exist to evade moral norms. By contrast, non-strategic mechanisms like motivated reasoning (for example), lead individuals automatically to bias their own perception of moral demands, leaving them vulnerable to others' judgments and their future self's self-assessment.

Finally, the third core element of our theory, and another key contribution to the literature, is its premise that *prosocial individuals*, who would otherwise serve as the main upholders of compliance, will be the most likely to engage in BSM, an assumption that we find supported in our results. Taken

together, the theory and evidence point to BSM's worrying potential to undermine and gradually erode cooperative behavior.

In prior work, we have shown that people utilize BSM also to overcome cognitive and motivational biases, such as regret and loss aversion, that can otherwise prevent them from advancing their self-interest. We found that decision-makers can reduce their regret-aversion-triggered endowment effect by sharing responsibility with an agent (Arlen & Tontrup, 2015a), or by following others' choices to limit the regret they anticipate to experience (Arlen & Tontrup, 2015b). We also demonstrate that people sophisticatedly exploit their own loss aversion as a commitment device to achieve desired outcomes (Tontrup & Sprigman, 2022). Together with our current work, these studies demonstrate the relevance of BSM theory across very different decision-making contexts.

Norm-shifting as a BSM Strategy. In this study, we test a particularly effective BSM strategy for decision-making in the fairness context: **norm-shifting**. Individuals engage in "norm-shifting" when they modify their decision-making environment so that the fairness norm applicable to their choice imposes lower demands.[2] For example, by investing effort, they may shift the applicable fairness norm from one that often demands that they share resources equally with others to one where it is fair to retain more than the other.

Norm-shifting is possible where social interactions are subject to multiple fairness norms, each placing different demands on decision-makers. Which norm becomes dominant depends on features of the decision-making environment—many of which decision-makers can actively influence through their behavior. For example, the perceived desert of the decision-maker (Oliver, 2024) , the deservingness of the interaction partner (Umer et al., 2022), and whether the decision-maker is interacting with someone for whom they are responsible or not, all can influence which and how fairness norms are applied.[3]

Many individuals encounter or practice norm-shifting in everyday life. Consider the allocation of household chores among roommates. The default norm is usually that each roommate contributes equally to tasks such as cleaning the kitchen. A roommate who wishes to avoid this duty, however, can take strategic steps to invoke an alternative fairness norm under which it would seem reasonable for them not to do so. For instance, they might assume another responsibility, such as cooking for the household. Or they might stop using dishes altogether by eating out. In each case, the actor alters the relevant decision-making environment so that a different fairness norm applies—one that justifies avoiding the chore.

Norm-shifting can be a particularly effective BSM strategy because the actions to render applicable a more favorable fairness norm can alter not only the decision-maker's own view of what fairness requires but also the perceptions of others. Thus, and crucially, norm-shifting does not merely allow the actor to *appear* fair or to convince themselves self-servingly that they are being fair

[2] Technically, what subjects act on directly in this BSM strategy is the decision-making environment. They seek to change it in ways that will cause themselves, and most others, to conclude that the equality fairness norm is not the appropriate norm to apply, but instead a different norm should guide judgment (here a merit-based norm). As changing the applicable norm is their ultimate goal in acting, we refer to this as "norm-shifting."

[3] In contrast, other factors do not affect which norm is applicable, but only make it easier for the decision-maker to disregard the applicable fairness norm. For instance, acting through an agent does not change the applicable norm or affect whether an allocation is fair or not; it only can reduce responsibility for an unfair allocative decision.

while actually pursuing self-interest; by reshaping which norm governs the situation, it renders the action fair under the newly applicable standard. In this way, norm-shifting enables individuals to act consistently with both self-interest and fairness, avoiding the self-image and social-image costs that would arise if they violated an applicable norm.

The Experiment. To test norm-shifting, we use a dictator game because it allows us to focus on distributive fairness norms. These norms are among the best documented and most studied social and ethical norms, with contours that are empirically well understood. Evidence from dictator games consistently shows that most individuals, when given an endowment to divide with another participant, will share with the other. They do so both to avoid the affront to their self-image as a good person and to avert the social-image costs of appearing unfair to others (Regner, 2021; Ariely et al., 2009; Andreoni & Bernheim, 2009).

In our dictator game implementation, an Allocator is given an endowment of €12 to divide between themselves and a Recipient—a protocol widely used in the literature that analyzes adherence to fairness norms. In the standard version of the game, the salient fairness norm is equal division, as confirmed by the overwhelming majority of our participants. In our norm-shifting treatments Allocators first receive the endowment in full and are then offered the option to complete a tedious real-effort task before making their allocation. Because completing the task does not increase the size of the endowment, it provides no additional benefit to the Allocator or the Recipient. We nevertheless expect a substantial number of Allocators to undertake the task because people generally agree that those who exert effort or perform work deserve more than those who do not (Hoffman & Spitzer, 1985; Sher, 1987; Lamont, 1994; Schmidtz, 2006). This design thus provides Allocators—but not Recipients—with an opportunity to claim desert and shift the applicable fairness norm: from equal division to an effort-based norm that justifies their self-interest in keeping a larger share of the endowment.[4] Accordingly, we hypothesize that Allocators who choose to complete the task will transfer less to Recipients than they would have in the standard dictator game.

To test whether Allocators are motivated by both self-image and social-image concerns, we implement two games that subjects complete consecutively: a Self-Image Game with no Observer and a Social-Image Game, which includes an Observer. In both games, we use a double-blind anonymity protocol, ensuring that only self-image considerations can play a role in the Self-Image Game. In the second game with an Observer, the Observer is informed about whether the Allocator worked and is also fully aware of the Allocator's transfer decisions. The Observer then can credibly express their judgment about whether the Allocator's transfer was fair. This design allows us to examine norm-shifting in the presence of both self-image and social-image costs.

Results: Norm-shifting is effective. Consistent with our BSM hypotheses, we find that between 30% and 40% of Allocators (depending on the treatment) chose to work in both the Self-Image and the Social-Image Game. As expected, completing the task shifted the applicable fairness norm for both Allocators and Observers. When we elicited their norm expectations, both groups concluded that Allocators were entitled to retain more if they had worked. Reflecting these expectations, working Allocators transferred significantly less to Recipients. While non-working Allocators

[4] Since the subjects receive the endowment independently of whether they take up the task or not, Allocators in our Treatments should have no reason to work other than to strategically shift to a more favorable effort-based norm.

transferred on average €5 of their €12 endowment, those who worked transferred only €2.4–€3.5—amounts consistent with what they and Observers judged to be fair under an effort-based norm. These transfers were also significantly lower than in the Baseline treatment, in which Allocators had no option to work.

Primarily prosocials use BSM. Our BSM hypothesis implies that Allocators should be more motivated to work the greater the burden that the fairness norm imposes on them. Because self- and social-image costs of an uneven allocation in the absence of work are higher for prosocial Allocators than for proself Allocators, we hypothesize that increasing prosociality should make Allocators more likely to undertake the work task. Purely proself subjects, who lack other-regarding concerns, can pursue self-interest without working since they should not experience significant self-image or social-image costs in our study when they act in their own favor. To test these predictions, we elicited each subject's social value orientation.

The results support our BSM hypotheses: prosocial Allocators are indeed more likely to work. Moreover, prosocials who work reduce their transfers more than proselfs do. Because prosocials face higher noncompliance costs, they give more in the absence of work and thus have greater scope to reduce transfers when they adjust to the effort-based norm. By contrast, proselfs already make self-favoring transfers without working, leaving them less room to reduce transfers if they do undertake the task.

Beyond prosociality, we also expect Allocators' strategic motivation to work to depend on their expectations about the fairness standard after working. Allocators should be more likely to work the greater their expectation of what they can fairly retain if they complete the task. Consistent with this prediction, we find that Allocators are more likely to take up the task the lower the transfer they believe will still be perceived as fair should they work.

Allocator Transfers Conform to the New Norm. Our BSM theory further suggests that Allocators who work can only secure the self- and social-image benefits of norm-shifting if their transfers fall within third parties' expectations of what constitutes a fair transfer by those who undertook the effort task. Supporting this hypothesis, we find that the vast majority of Allocators adjust their transfers to match their own fairness expectations, selecting amounts they believe the Observer will regard as fair—indicating that they aim to protect both self-image and social-image. Indeed, on average, their transfers are judged as fair by two-thirds of Observers.

Interestingly, all these results hold in both the Social-Image and the Self-Image Game: Allocators still chose to work even when they remained completely anonymous and no one is observing their work decision in the Self-Image Game.

Allocators also appear to work in order to give themselves "moral wiggle room" to retain more of the endowment. By working, they shift the applicable standard from an equal-division norm to an effort-based norm—one that both they and others perceive as broader and more ambiguous. This ambiguity creates space for Allocators to transfer less than the amount they are fully confident is fair, provided the transfer still falls within the range that Observers regard as acceptable. To test this, we elicited Observers' beliefs about the effort-based norm and found that they were indeed more uncertain about what constituted a fair transfer compared to the clear equality norm. Allocators

appear to anticipate this uncertainty, placing their transfers near the lower bound of what they expect Observers to view as fair.

Using BSM can be socially costly. Norm-shifting—even when the new norm is socially accepted—can still be morally costly. Allocators incur such costs when Observers know, and recognize the strategic motivation that underlies the Allocators' decision to work, as opposed to merely knowing that the Allocator worked. To test this, we conducted two treatments: in Treatment 1, Observers of Allocators who worked were informed only that the Allocator had worked, without being told that the Allocator had chosen to work; by contrast, in Treatment 2, Observers were explicitly informed that Allocators had chosen to work. When Observers learned that Allocators chose to work, they attributed less merit to the work, apparently because they understood the work choice was strategic. Anticipating this reaction, Allocators responded with higher average transfers in Treatment 2 compared with Treatment 1. Yet, nevertheless, Allocators who worked could and did fairly transfer significantly less than those who did not.

In sum, our experiments support the conclusion that individuals can and do reshape their normative decision-making environment to secure personal advantage, and that norm-shifting enables them to do so largely without incurring the disutility of self-image or social-image costs, in particular when the BSM strategy goes unnoticed.

Implications for legal compliance. This evidence is important for several reasons. First, our study shows that BSM applies beyond the area of cognitive biases to decisions implicating social, legal, and moral norm compliance. Second, this study complements our other projects on BSM in the fairness domain: delegation, which examines how individuals can strategically dilute their accountability by involving artificial intelligence in their decisions (Tontrup & Sprigman 2025), and accessing information about others' compliance, which allows individuals to reduce their moral burden by comparing themselves with others (work in progress). Taken together, these studies demonstrate how individuals actively reshape their decision-making environment to mute the demands of social norms. Consequently, identifying, testing, and assessing the possible real-world application of additional BSM strategies remains an important task for future research.

Our results have policy implications because they show that individuals can strategically avoid the demands of other-regarding social norms at relatively low cost. In particular, prosocial types—those who would otherwise uphold collective adherence to social norms— are most likely to engage in BSM, which may contribute to the erosion of compliance, a troubling finding given that norm stability often depends on prosocial actors.

Through the interplay between social and legal norms, norm-shifting can also undermine legal compliance. For example, legal rules often rest on social norms that protect at least partially the same interests as the law. If people can change the decision-making environment such that the underlying social norm appears satisfied, they can make legal violations less costly to their self- and social-image. Consider a restaurant pet ban, which protects patrons' health, safety, and comfort. Pet owners violating the law by entering restaurants with their pet can shift the normative judgment of their conduct by encouraging other customers to pet and play with their dog, providing salient evidence that customers do not perceive the dog as a safety threat or source of discomfort. This appears to satisfy the social norm against inflicting harm on others: no visible harm seems to be caused with the law only being formally violated. The pet owner may even receive social approval,

and any customer wanting to insist on the law being enforced may anticipate social-image costs for doing so.

BSM can also be used to establish social norms that counter the purpose of legal norms. Consider laws that prohibit harmful conduct by employees of corporations. Such legal injunctions against bribery for example are most effective when they leverage employees' preferences to comply with social norms, as often the risk of sanction and punishment is low. Yet, the social norm that is most salient for employees is often defined by the conduct of those immediately around them. As a result, managers whose bonuses depend on business success can undermine the law's expressive effect by informing employees that competitors routinely pay bribes and that the company's survival depends on doing the same. This narrative activates social norms of team loyalty, in-group solidarity, and competitive fairness. These norms make violating anti-bribery laws relatively less costly, because legal compliance would entail letting down one's colleagues. Bribery is recast as an act of organizational duty, offering self-image protection and social-image protection (Arlen & Kornhauser, 2023).

BSM thus can operate in legal contexts, enabling individuals to shift or invoke social norms so that legal violations appear normatively justified. Our study also shows that the social costs of using such strategies can be particularly low when actors can conceal their self-interested use of BSM by framing it in terms of cooperative norms—such as "respecting others' choices" in the smoking example or "being a loyal team player" in the bribery example.

In Section II, we review the literature on fairness norms, specify the theory that underlies our research, and differentiate BSM from other behavioral devices that have been previously studied. In Section III, we explain the design of our experiment. In Section IV, we detail our hypotheses and report the experimental results. In Section V, we analyze the policy implications of our findings.

II. Theory, Literature and Predictions.

In this section, we examine (A) what motivates individuals to employ Behavioral Self-Management (BSM) strategies and (B) why they are able to do so. We then focus on (C) the specific BSM strategy investigated in this study—norm-shifting—and contrast it with other cognitive mechanisms and forms of deliberate moral behavior that also serve to reduce the self- and social-image costs of acting selfishly. This comparison helps to clarify the distinctive nature of norm-shifting and underscores the specific contribution of our study.

A. Prosocial Behavior and Compliance with (Fairness) Norms

A large body of evidence shows that people's decisions are motivated by both self-interest and a willingness to adhere to fairness norms (Lin et al., 2020; Hoffman & Spitzer, 1996; Engel, 2011; see also Fischbacher et al., 2001; Johnson & Mislin, 2011). People want to perceive themselves and be perceived by others as fair and prosocial to maintain their self- and social-image (Vanberg, 2008; Andreoni & Bernheim, 2009; Ariely et al., 2009; Grossman, 2015; Grossman & van der Weele, 2017; Mischkowski et al., 2019). The disutility to their self- and social-image that people experience from actions that they perceive as unfair or expect others to conclude are unfair motivates them to correspond to fairness norms, even when they would benefit financially from not doing so. Similarly, people exhibit a preference for complying with other-regarding social norms. In the field, we see

many people donating blood (Götte et al., 2010), getting vaccinated even though their personal risk of illness is low (Aldridge et al., 2024), and doing volunteer work and donating to charity even when remaining anonymous (Andreoni, 1990). This evidence would appear to offer solace to legal policy makers and the leaders of organizations who hope that they can rely on people's robust other-regarding preferences to achieve compliance with ethical and legal standards.

People's motivation to comply with fairness norms is shaped by a range of factors, one of the most important being their personal degree of prosociality, which we measure directly in this study. A robust body of research demonstrates that individuals with stronger prosocial preferences are more likely to comply with social and fairness norms (Fehr et al., 2004; van Lange, 1999; Capraro & Rand, 2018; Capraro, 2018). For such individuals, norm violations are psychologically costly. The violations threaten their self-image and social-image, and they are more likely to experience negative emotions, such as guilt and shame (Regner, 2021; Tangney et al., 2007). Thus, prosocial individuals have the most to gain from using BSM strategies to reduce their social-image and self-image costs when violating a norm.

By contrast, individuals with a proself orientation should be less sensitive to fairness norms as they care less about the outcomes of others and whether these are fair or not. Their desire to be viewed as acting in accordance with fairness norms is also lower. Therefore, proselfs who can be expected to give little even without using BSM, have little incentive to engage in BSM strategies.

B. Fairness Norms are Context Specific and Can be Altered

In the fairness context, a key prerequisite for using BSM and norm-shifting is that both the content and applicability of fairness norms are shaped by social factors that decision-makers can influence. This mutability is reflected in many behavioral fairness experiments.

Evidence from standard dictator and ultimatum games shows that the salient default social norm in these settings is one of equal distribution of the allocation. Inequality aversion explains the strong salience of this equal distribution norm (Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000; Andreoni & Millner, 2002). This norm has been shown to be shared by Allocators, Recipients, and Observers (Engel, 2011; List, 2007).

Yet, a norm of equal distribution is not the only fairness norm that may apply to a dictator game (see List 2007; Arlen 1998). The fairness norm that applies can depend on many elements of the decision-making context, some of which can produce a consensus that either the Allocator or the Recipient deserves more. These factors include whether the Allocator earned the endowment through skill or effort (Engel, 2011; Oliver, 2024), whether the Recipient is especially deserving (e.g., is a charitable organization; Eckel & Grossman, 1996), and the social distance between Allocator and Recipient (Hoffman et al., 1996; Frey & Bohnet, 1999; Lazear et al., 2012), among other considerations.[5] The decision-maker can control some of these factors. In our study for example, Allocators invest effort to increase their own desert and shift the norm in their favor.

However, not all ways of changing the decision-making environment are equally effective. When an Allocator *works* to shift the norm, as in our study, both Allocators and others tend to believe that the Allocator deserves a larger share; this effort-based norm has a long history in moral

[5]Other factors that affect the generosity of transfers are whether the Recipients are known identified individuals or groups who are socially or geographically proximate versus they are socially distant "statistical" people (Hoffman et al., 1996).

reasoning and appears to be generally accepted (see Faillo et al., 2019). In contrast, if the decision-maker delegates the decision to an agent, they change the normative environment to lower their perceived responsibility (Tontrup & Sprigman, 2025). Likewise, when individuals strategically ignore information that would increase their moral costs (Dana et al., 2007; Mommsen Ohndorf 2020; Offer et al., 2024), they aim to manipulate their perceived responsibility for an unfair outcome. However, they do not shift the applicable fairness norm. While both are BSM strategies, the effectiveness of these strategies is an empirical question we aim to answer.

For BSM to be employed, it is not enough that the decision environment **can** be altered to reduce the decision-maker's costs of norm compliance—decision-makers must also be aware of these opportunities to change their normative decision-making environment. The nature of fairness preferences seems to assure that they are not only strong, but also develop early. Blake et al. (2014) and Shaw and Olson (2012) show that children aged six to eight will incur costs, such as discarding a resource, to prevent unequal distributions. By the age of three, children already demonstrate an understanding of multiple fairness principles—most notably equality and merit—and recognize that a fair allocation may require giving more to those who have worked harder or who have fewer resources (Starmans et al. 2017; Blake et al., 2014; Kanngiesser & Warneken, 2012) but also—as in our study—that factors like investing effort may justify an unequal distribution in their favor (Faillo et al., 2019; Bicchieri, 2017).

Accordingly, the literature suggests that people understand very early in their lives from experience that fairness norms can vary across social contexts and that they and others can take actions that will alter the applicable fairness norm. We test whether individuals use this experience strategically to avoid the disutility of damage to their self-image or social-image they would suffer if they directly violate the fairness norm rather than invoking a less demanding effort-based norm through their effort and then complying with that norm.

C. Norm-Shifting and What it Contributes to the Literature on Norm-Compliance

We posit that BSM via norm-shifting has three distinctive elements: (A) individuals deliberately alter their decision-making environment to obtain a benefit, (B) while preserving their self- and social-image, and (C) they do so strategically—as choice architects. Norm-shifting is effective because it *objectively* changes the normative context of the decision. By this, we mean that third parties will judge the same behavior differently—and more favorably—from a normative perspective after the decision-maker changes the decision-making context. Norm-shifting through effort illustrates this: prior to the norm shift, a transfer below the equal split is perceived by most Observers (and Allocators) as unfair. However, once effort has been exerted, the same transfer is viewed as fair, because it is now evaluated relative to an effort-based norm.

This norm-shift sets our work apart from prior studies that show how people manipulate their own perception of what constitutes fair behavior without actually changing the normative decision environment. Examples that illustrate this difference include cognitive mechanisms such as self-serving bias or motivated reasoning, in which individuals persuade themselves that the norm most advantageous to them is also the most fair—such as interpreting fairness to mean proportional rather than equal tax contributions (Balliet et al., 2009). Another example is moral disengagement, where individuals justify norm violations by questioning the moral worth of those affected (Cain et

al., 2005; Hofmann et al., 2014; Bandura, 1999). In both cases, the individual reinterprets moral obligations internally, but the objective structure of the normative decision environment—and the applicable fairness or social norm—remains unchanged.

Other examples are studies on *moral wiggle room*, which examine how people exploit normative ambiguity within a preexisting social context. Consider Dana et al. 's study from 2007: individuals are presented with a binary dictator game in which the consequences for the Recipient are initially concealed. Dictators can choose to remain ignorant about the outcomes their decisions impose on the Recipient and opt for the more selfish allocation. While this allows them to avoid directly acknowledging the harm caused, and thus mitigating their self-image costs, it manipulates only their own perception of the moral implications of their choice. Crucially, they do not change the objective decision-making environment; they merely obscure the recipients' outcomes from themselves (unless they could withhold the information from others too).

In the same way, moral licensing—where individuals use past and unrelated good deeds to justify later actions that conflict with moral norms (Mazar & Zhong, 2010)—does not change whether the subsequent behavior is morally acceptable. Instead, it allows individuals to manipulate their own moral judgment by referencing unrelated past behavior to justify their current morally questionable actions.

Only a few studies have examined strategic behavior that aims to change the decision-making environment. In Spiekermann and Weiss (2016), for example, individuals strategically acquire only information that supports a self-serving application of norms. Like the mechanisms discussed above, this again produces a primarily internal manipulation: individuals strategically steer their beliefs about what fairness requires by creating a biased representation of the true state of the world—which, in Spiekermann and Weiss's case, is salient to third parties.

One consequence of not altering or biasing the normative decision environment is that self-image, and especially social-image costs, cannot be as effectively reduced. In Dana et al. (2007), if the Allocator chooses blindly an unfair allocation, this will generate self- and social-image costs. Observers will reasonably infer that the decision-maker disregarded the Recipient's outcome and chose ignorance for self-serving reasons. Thus, strategic ignorance cannot shield the decision-maker from moral costs as effectively as a socially accepted norm shift that provides a normative reason to view the self-interested behavior of the decision-maker more favorably.

The same applies to Spiekermann and Weiss. Unless such selective information acquisition is concealed, third party observers can infer not only the self-serving intent but also understand that the application of the norm rests on a misrepresentation of the true state of the world and the outcome might thus be incorrect.

As we have seen above, delegation to agents is different in that it actually changes the decision-making environment in a normative way that reduces the Principals' self- and social-image costs. Principals will be attributed less responsibility also by impartial third-parties if the agent carries out the unfair allocation (Gerstenberg & Lagnado, 2012). Hamman, Loewenstein and Weber (2010) found that Allocators in a dictator game will choose to delegate the allocation choice to those agents who had previously transferred less to Recipients than other agents, suggesting they anticipated the agent's self-serving behavior and used it to their own advantage. They also found that

Allocators who made unfair transfers through an agent incurred lower social costs from third-parties than when they did so directly (see also Bartling & Fischbacher, 2012; Grossman & Oexl, 2012; Hill, 2015). However, there is an important distinction: unlike norm-shifting, delegation does not eliminate the norm violation itself; it merely reduces the principal's perceived responsibility for the norm violation. Nevertheless, by reducing accountability, delegation—just as with norm-shifting—provides a normative reason to view the decision-maker's self-interested behavior more favorably. Insofar as delegation and norm-shifting both protect the decision-maker's self-image and social-image, they are potentially equally effective BSM strategies, even though they invoke very different normative justifications.

A second important element of BSM is that decision-makers act as choice architects of their own decision-making environment. Rather than passively accepting existing fairness norms, they strategically redesign the normative environment by shifting how their behavior is evaluated. Having reshaped the normative standards, they then exploit the new environment by lowering transfers while maintaining their self- and social-image—keeping their behavior within what they expect others might still regard as fair.

Effort-based norm-shifting exemplifies this form of choice architecture: Allocators actively engineer a shift from an equal-split norm to an effort-based norm, which generates greater uncertainty and heterogeneity in fairness judgments. This expanded normative space allows individuals to justify lower transfers that may plausibly fall within the range of what observers view as fair. Our results show that individuals appear to calibrate their transfers to meet such a range—trading off personal gain against anticipated image costs. Thus, as choice architects of their normative decision-making environment, individuals both create wiggle room and strategically exploit it.

In contrast to this choice-architect dimension of BSM and norm-shifting, in the existing wiggle room studies we are aware of, individuals typically exploit *preexisting* ambiguity—they do not *create* ambiguity in order to benefit from it. By acting as choice architects through actions that change the normative structure of the situation, individuals can make norm-conforming behavior more compatible with their self-interest.[6] Now we turn to our experiment.

III. Experimental Design

A. Self-Image and Social-Image Games

We use the dictator game as the basic structure of our study. The core game has two players, the Allocator and the Recipient. Roles are randomly assigned, and players are randomly matched to each other. The Allocator is given an endowment of €12 and can decide how much of this endowment to keep or transfer to the Recipient.

[6] While the choice-architect perspective is common in the business literature, this literature examines the multiple ways in which businesses can strategically influence employees' decision-making environment to alter the likelihood that they will comply with ethical or legal rules. It focuses on how companies may, sometimes unintentionally, structure their employees' working environment in ways that promote misconduct, and the steps that companies can take to foster compliance (see generally, Arlen & Kornhauser, 2023; Bazerman & Tenbrunsel, 2011); Langevoort, 2018; Feldman, 2018; Feldman & Kaplan, 2021). BSM by contrast, focuses on actions employees or others may take to restructure their *own* decision-making environment to mute the demands of injunctive ethical or legal norms.

Each subject in our main experiment plays two variations of the dictator game—the Self-Image Game and the Social-Image Game. To keep incentives constant between these two games—for example, to avoid income effects—we use the strategy method: subjects are paid only once, either for the first or the second game. They are not told which game they will be paid for until after the experiment. The Allocators receive as payment the portion of the €12 endowment they keep for themselves. The Recipient receives whatever amount the Allocator assigns to her.

The Self-Image Game allows us to isolate self-image motivations to work as Allocators' work decisions are private and not subject to assessment by Observers. A double-blind procedure ensures that neither the experimenters nor the Recipients can identify the Allocators (Hoffman et al. 1994, identifying experimenter effects). Notably, Recipients are not aware that Allocators are offered a work option and accordingly, they do not learn whether Allocators worked or not. In constructing these double-blind procedures in the Self-Image Game, we eliminated the possibility that Allocators might choose to work due to concerns about their social-image. If Allocators still choose to complete the work task in the Self-Image Game, it suggests they are driven purely by self-image concerns—working in order to justify their transfer decisions to themselves and to appear more deserving in their own eyes.

The Social-Image Game differs from the Self-Image Game because it includes an Observer. In the Social-Image Game, the Allocator knows that his or her work also affects the *Observers'* fairness judgments. As a consequence, an Allocator in the Social-Image Game who fails to comply with the fairness norm will bear social-image costs related to the Observer's judgment of the Allocator's actions. In addition, in the Social-Image Game the Recipient knows that the Allocator worked if the particular Allocator with which a Recipient is paired chose to work. In contrast, in the Self-Image Game the Recipient is not informed that the Allocator worked. Thus, only work affects the Allocators' own perception of the fairness of their choices. The Observer is given an endowment and asked to judge what a fair Allocator should transfer to the Recipient. The Observer is instructed to use the 50€-cent endowment she or he is given to reward a transfer he or she perceives to be fair. The Observer can give the Allocator all, half, or none of their endowment. Since we want to interpret the Observer's reward as an expression of the Observer's approval of the Allocator's transfer decision, the Observer cannot transfer any part of her or his endowment to the Recipient, but only to the Allocator. Anything not transferred to the Allocator is automatically returned to the experimenter. While the Allocator has full knowledge of the Observer making a judgment, the prospect of receiving a reward from the Observer does not materially change the Allocator's monetary incentives, as the maximum award is 50 €-cents and so only half of the extra euro (at minimum) the Allocator could gain by lowering their transfer decision.

B. Treatments, Baseline and Control

We implemented a Baseline condition, two treatments and a Control; all included both a Self-Image and a Social-Image Game.

1. Baseline

The Baseline implements the standard dictator game—without a work option and without an Observer— as a benchmark for our treatments.

2. Treatments

In each of our treatments, Allocators were offered the choice to complete the work task, which entailed counting how many times a particular number appears in a 10x20 matrix. The work task is tedious, uninteresting, and not useful to the experimenters or anyone else. The only plausible reason for Allocators to undertake the work task was to change the decision-making environment to one where the Allocator had worked while the Recipient had not, thereby triggering an effort-based norm under which it is fair for the Allocator to retain more than the equal distribution. Thus, completing the task was motivated both by the self-interested pursuit of higher payoffs and by the desire to comply with norms, even if only the less demanding merit norm. We first describe the two Treatments in the Social-Image Game.

In Treatment 1, the Observer learned whether the Allocator worked and the amount the Allocator transferred. However, Observers did not know that the decision to work was voluntary and that the Allocator's endowment was unaffected by that decision. To confirm Observers' understanding of the instructions, we asked them to indicate what they believed motivated Allocators to work. The results show that a large majority assumed that Allocators following their instructions had to perform the task in order to complete the experiment, while those who did not work were simply not instructed to do so. Only a few of the Observers thought that the Allocators had been given a choice whether to work. This manipulation check confirms that in Treatment 1, Allocators can deploy work as a concealed BSM strategy, knowing that Observers can consider the effort in their judgment but not its strategic BSM purpose.

Treatment 2 in the Social-Image Game is identical to Treatment 1 except that Observers were explicitly informed that Allocators were given a choice whether to work. Thus, any decision by Allocators to complete the work task could, in the eyes of Treatment 2's informed Observer, be motivated by the Allocator's desire to make themselves appear more deserving of a larger share of the windfall endowment—i.e., to strategically effectuate a norm shift.

Social-image costs could affect both the Allocator's transfer decision and the Allocator's decision to work. Comparing the transfers in Treatment 1 and Treatment 2 allows us to examine whether Allocators face social-image costs should they be observed engaging in BSM (i.e., choosing to work in order to shift to a fairness norm more favorable to them). We assume Allocators will anticipate that Observers will set the fairness norm for the required transfer lower in Treatment 1 (where at least some Observers falsely believe that Allocators did not have a choice whether to work or not) compared to Treatment 2 (where Observers understand that the Allocators decided to work for self-interested reasons). Accordingly, we expect Allocators to transfer less in Treatment 1, where the Observer is unaware that the decision to work is self-interested, than in Treatment 2, where the Observer is fully aware of that fact. Since Allocators themselves know in both treatments that they chose to work for self-interested reasons, the two treatments hold self-image costs constant, such that any treatment effect can be attributed to differences in social-image costs attributable to whether the Observer knows the Allocator's work choice was strategic.

In the Self-Image Game we have only one Treatment. Since the Self-Image Game does not involve an Observer—and the only distinction between Treatment 1 and Treatment 2 concerns what the Observer knows about the Allocator's work option— the two treatments collapse into one treatment condition in the Self-Image Game.

The Self-Image Game enables us to test whether people undertake work purely to ameliorate the self-image costs of retaining more than half of the allocation. The comparison between the Social-Image Game and the Self-Image Game enables us to examine the impact of introducing social-image costs on both the choice to engage in norm-shifting and the transfer choice.

3. Control

As a Control we conduct a treatment designed to test whether the decision to work might be driven by motivations other than BSM. For example, Allocators might prefer not to receive the endowment as a mere windfall but instead wish to work for it. They may also work out of a sense of gratitude toward the experimenters, out of curiosity about the task, or due to demand effects (i.e., believing that working might support the researchers in their research).

As explained earlier, we designed our work task to rule out these potential confounds (see the internal validity section). Nevertheless, we conducted a Control treatment as it allows us to directly test for and rule out these alternative explanations. As with all other treatments the Control consists of the two games: the Self-Image Game and the Social-Image Game. However, Allocators receive the €12 endowment as a lump sum that they cannot distribute between themselves and the Recipient. Instead, Allocators are given virtual points of no monetary value to distribute between the Recipient and themselves in the dictator game.

Our BSM theory suggests that since Allocators transfer nonmonetary tokens rather than money, they will transfer six tokens consistent with the equal-distribution norm (or even more) as a costless way to maintain their social-image and self-image. Consequently, according to our theory they lack any incentive to work: if the tokens have no monetary value, they have no motivation to work to transfer fewer tokens. In contrast, if the decision to work was driven by curiosity, or a desire to help the experimenters with their research or by a desire to earn the endowment rather than receive it as a windfall, we nevertheless should see Allocators work in the Control group with a similar frequency as in the treatments of the Social-Image Game and Self-Image Game. Conversely, if Allocators are more likely to work in the Treatments than in Control, we can conclude that the difference is driven by a motivation to use BSM.

Table 1. *Overview of Treatments & Games*

Treatment	What is allocated	Work Option	What the Observer Knows in Social-Image Game
Baseline	€12 endowment to allocate	No	Knows there is no work option offered
Control	Non-monetary tokens to allocate (€12 paid as a lump sum to keep)	Yes	Knows Allocators can choose to work & whether they did work
Treatment 1	€12 endowment to allocate	Yes	Knows whether Allocator did work; does **not** know Allocator can choose to work
Treatment 2	€12 endowment to allocate	Yes	Knows Allocators can choose to work & whether they did work

3. The Work Task

The work task consists of a 10 x 20 matrix which displays numbers between 1 and 9. Subjects who elected to complete the task had to count how often a specified digit appeared in the table. Participants who answered incorrectly had to wait 15 seconds before making a new attempt. The buffer between entries makes guessing less time-efficient. On average, completing the task took approximately 2 minutes. To ensure that subjects focus on the experimental task, we implement a time limit. Participants are informed that if they fail to make an input after 180 seconds or if they log out, we exclude them from the experiment without receiving payment. A figure showing the task is in the Appendix.

To isolate the BSM motivation for working, we selected an uninteresting number-counting task whose completion did not directly benefit the Allocator (other than as a BSM strategy), the Recipient, the experimenter, or society. We chose a pure-effort task in order to prevent intrinsic or prosocial motivation, curiosity or demand effects from confounding our BSM results. We expected, correctly,[7] that Allocators and Observers would conclude that Allocators who exerted effort could retain more than those who do not, relying on evidence that people believe that those who exerted effort should be rewarded, and expect others to share this belief (Ogawa et al., 2010; see Cherry et al., 2002). This merit-based fairness norm appears to be based primarily on the belief that work or effort confers merit and that merit deserves a greater reward, even when the work in itself is not socially beneficial (Bhattacharya & Mollerstrom, 2025; Andre, 2025). Studies find that people believe effort deserves greater reward even when the effort neither increased the allocation available to the Allocator and Recipient nor benefited society (Ogawa et al., 2010). Moreover, small increases in merit have been shown to justify large differences in rewards (Bhattacharya & Mollerstrom, 2025; Cappelen & de Haan, 2023)—differences exist even when access to work opportunities resulted from pure chance (Bhattacharya & Mollerstrom, 2025; Cappelen & de Haan, 2023).[8]

The task was thoroughly explained, and the Allocators were given a trial to complete the task before making their decision. Importantly, Allocators were neither directly nor indirectly compensated for completing the task, nor did Recipients receive any additional payment if the task was completed. Therefore, the only plausible reason for Allocators to undertake the work task was self-interest: specifically, to change the decision-making environment to one where the Allocator had worked while the Recipient had not, thereby rendering applicable an effort-based fairness norm that would justify transferring less than the €6, which represented the equal distribution.

C. Eliciting Norm Expectations

1. Observers' Norm Expectations in the Social-Image Game

Observers were instructed to identify the fairness norm in the dictator game with and without work. Observers were asked what an Allocator should transfer to the Recipient in order to act fairly, first assuming that the Allocator completed the work and then assuming that the Allocator did not work. The design provides us with between-subjects and within-subject data to assess the magnitude of the norm shift caused by the decision to work. That is, by comparing Observers' perceptions of the

[7] See infra text accompanying notes X-X.

[8] Indeed, people who work hard are deemed to be more moral or deserving, relative to those who engage in little or no effort (Harris & Fiske, 2011; Petersen et al., 2012), even when the effort produces little or no material value. (Celniker et al., 2023).

fair transfer when work is completed with Observers' perceptions in *Baseline*, when no work option is offered, we have between-subjects evidence that work shifts the norm downward in the eyes of the impartial observers. We also have within-subject evidence, as the difference between the two assessments for an individual Observer reveals whether that Observer thinks that the fairness norm is shifted by the completion of the work task.

To elicit the values, we asked each Observer, for each possible transfer from 0 up to €12, whether that transfer would be in accordance with their perception of the fairness norm. Observers were presented with three levels of confidence to choose among: 100% (completely confident), 50% (somewhat confident), and 0% (not confident). Each Observer received a €0.50 endowment and was instructed to reward the full €0.50 if they perceive the transfer amount as clearly within the fairness norm, to award €0.25 if they are only somewhat confident, and to award nothing if they have no confidence that the transfer still falls inside the fairness norm. Any amount the Observer did not give to the Allocator was automatically returned to the experimenter.

2. Allocators' Norm Expectations in Social-Image Game

We also elicited Allocators' expectations about what transfer amount the Observer will consider fair. To ensure that belief elicitation does not influence transfer decisions, Allocators were queried after they made their transfer choice. In turn, to assess whether the belief elicitation is biased by subjects' own transfer choices, we compare their stated beliefs to those elicited in a pilot session where participants did not make any prior transfer decisions.

Allocators assessed each possible transfer amount (€0–€12) and indicated their confidence that the Observer would judge it Fair: 100% (completely confident), 50% (somewhat confident), or 0% (not confident).

To incentivize Allocators' assessment, each Allocator received a separate €1 endowment The computer randomly drew a stake from this €1 endowment to be wagered on the Allocator's assessment; the Allocator kept the remainder. The computer then randomly selected one of the 13 transfer amounts the Allocator had evaluated and compared the Allocator's stated confidence for that amount with the Observer's actual fairness judgment. Payoffs depended on both the stated confidence level and the accuracy of the prediction: if the Allocator indicated to have 100% confidence, that the Observer would consider the particular transfer fair the stake was tripled if the Observer indeed considered the transfer fair and lost otherwise; if the Allocator indicated to have 50% confidence, the stake was multiplied by 2.5 if correct and halved if the Allocator was wrong; at 0% confidence, the stake was tripled if correct—that is, if the Observer indeed did not judge the transfer fair—and lost if incorrect.

The payment structure incentivizes honest reporting. This is intuitive for cases such as a €12 transfer, which Allocators are likely to correctly believe the Observer will judge as fair, or a €0 transfer, which they are likely to correctly believe the Observer will view as unfair. Thus, choosing 50% when they in fact hold strong (in their view likely correct) beliefs lowers expected payoffs. Moreover, because only a randomly drawn portion of the €1 is staked, Allocators can expect to likely keep some money even when a bet is lost, which attenuates the motivation to choose 50% out of risk aversion.

This measure allows us to identify a range of transfer amounts where Allocators are uncertain (only somewhat confident) that a transfer satisfies the Observer's norm expectation. When an Allocator makes a transfer within this uncertainty range, we can infer that they likely intended to exploit ambiguity. It also shows how far Allocators believe they have shifted the Observer's fairness norm from equal division.

3. Allocators' (Hypothetical) Norm Expectations in Self-Image Game

We elicited what Allocators expect third parties would consider a fair transfer conditioned on whether they have worked or not, to test if Allocators are motivated in their work and transfer decisions by this expectation. Since these were beliefs about *hypothetical* third parties, they could not be incentivized.

D. Social Value Orientation

We elicited subjects' social value orientation (SVO), a measure of how much an individual cares about other players' outcome in relation to their own, using the ring measure developed by van Dijk, Sonnemans and van Winden (2002). The measure consists of 32 binary dictator games in which subjects decide how much of their own payoff to sacrifice for increases (or decreases) in others' payoffs. Higher values indicate stronger other-regarding preferences and prosociality. In the Appendix, we describe in detail how the measure is constructed. The ring measure has been shown to be predictive of cooperative behavior in many studies and across different games and scenarios: for example, in public goods games studied in De Cremer and van Lange (2001) and Fiedler et al. (2012), in one-shot and repeated prisoner's dilemma tasks in Balliet et al. (2009) and Pletzer et al. (2013), in investment games in Kanagaretnam et al. (2009), in trust games in Yamagishi et al. (2013) and Haesevoets et al. (2014), and in field studies by van Lange et al. (2007). Hollander-Blumoff (2017) analyzes how prosociality influences legal compliance.

E. Control Questions

We asked our subjects a set of control questions to ensure their comprehension of the study's instructions. For example, we inquired whether they can increase their endowment by completing the work task. The complete list of questions can be found in the Appendix.

F. Methods

1. Recruitment

We recruited n=100 subjects for each Treatment: Baseline, Treatment 1, Treatment 2 and Control. Once the target *N* was reached the experiment was stopped. The participants were either current students or graduates of the University of Münster in Germany. Participants are studying or have studied in a number of different academic departments. Approximately 23% of the sample had graduated and were employed outside of the university. Using LimeSurvey, we sent participants an invitation email with a link to the study. The link became inactive once used. The invitation informed participants of the amount of time they would need to complete the experiment, in order to minimize the number of subjects who discontinue participation.

2. Demographics

We obtained age and gender information as both can influence prosociality. The gender composition was roughly balanced, with slightly more women than men. Age showed little variance across participants. Women transferred nonsignificantly more to Recipients than men across treatments, while age showed no relationship with transfer amounts. As neither demographic variable significantly affected results, we do not report them in our regression analysis.

3. Double Anonymity

To ensure that participants are protected by double-blind anonymity throughout the experiment and payment.[9] We advised subjects to create and use an email address without any personal identifiers that is also not linked to their persona and activities on the internet. Where required we also provided subjects with assistance in creating such an email address for the purpose of the study. If subjects logged in with an email address that contained identifiers, we asked them to participate with a different address. The entries with identifiers were deleted. Likewise subjects also did not learn the identity of the researchers until after the experiment was completed. We followed this double-blind protocol in order to be able to cleanly isolate self-image concerns as a motivation for using BSM, and to minimize potential demand effects.

IV. Hypothesis and Results

We aim to separately analyze each element of our Behavioral Self-Management (BSM) theory, with the presentation of our results following the structure of our theory's framework. According to our BSM theory, individuals will **(A)** actively change their decision environment by choosing to perform the work task. They do so in order to **(B)** gain self-interested benefits by reducing their transfers while **(C)** protecting their self-image and social-image through shifting the applicable fairness norm from equal distribution to one based on effort. They seek and do take advantage of both the objective norm shift and the "moral wiggle room" created by the shift to a broader norm with more uncertain boundaries. Finally, they engage in this behavior **(D)** in a strategic and behaviorally rational manner; that is, prosocials who benefit more from shifting the norm are also more likely to work.

We begin our BSM analysis by testing the first prediction: Allocators will actively seek to change their decision environment by completing the work task.

A. Changing the Decision-Environment: Allocators Choose to Work

Our theory predicts that Allocators have an incentive to work in both the Social-Image Game and the Self-Image Game in order to shift the fairness norm. Under our theory, some Allocators will not work. Some may not work because they are truly self-interested and do not incur internal costs from anticipated loss of self-image or social-image when they fail to act fairly to others (see infra Section D). Others might not do so if they conclude that the effort-based norm does not permit a sufficiently high increase in transfer to outweigh their cost from working (see infra Section C).

We predict that Allocators who work will do so in order to change the decision-making environment to one where the applicable norm is an effort-based norm, under which it is fair for the Allocator to retain more than half of the distribution.

[9] For a discussion of why this is important see Hoffman et. al (1996).

To establish that BSM is driving the work decision of those who choose to work, as opposed to potential demand or desirability effects for working, we compare the frequency with which Allocators choose to work in both our Self-Image and our Social-Image Game with a Control condition. In the Control condition, Allocators receive the €12 as a lump sum to keep. They are given the choice to transfer tokens of no monetary value to a Recipient. Prior to making that choice, they are given an option to undertake the work task. If demand effects, desirability effects, or curiosity were driving their work decisions, we should see no difference between our treatments and the Control. However, if we observe a treatment effect, it must be driven by Allocators using BSM.

Hypothesis A_1: ***A significantly larger proportion of Allocators will choose to work in the Treatments than in the Control condition.***

In the Social-Image Game, 40% of Allocators worked in Treatment 1 and 30% worked in Treatment 2. By contrast, only 10% worked in the Control condition. Two-proportion z-tests confirm that the take-up rate is significantly higher in both Treatment 1 (z= 7.07, p< 0.01) and Treatment 2 (z= 5.94, p< 0.01) than in Control, supporting $\mathbf{A_1}$.[10]

In the Self-Image Game, 22% (n=44) of Allocators in the Treatment condition choose to work, whereas only 7% do so in the Self-Image Game of the Control group. A two-proportion z-test confirms that this difference is significant (z= 3.589, p< 0.01), supporting $\mathbf{A_1}$ and showing that participants actively engage in BSM even when they are driven solely by self-image concerns.

To confirm our analysis, we run a separate logistic regression for the Self-Image Game and the Social-Image Game using the Control in both as the reference category. We include dummy variables for the Treatments. We use a separate treatment dummy for the Self-Image Game because it differs conceptually from Treatments 1 and 2 in the Social-Image Game: the Self-Image Game has no Observer, whereas both Social-Image treatments include an Observer and vary in what the Observer knows. The logistic regression confirms $\mathbf{A_1}$: the odds that an Allocator chooses to work are significantly higher in the Treatments than the Control (see Table 2 below).

[10] Comparing Treatment 1 with Treatment 2 fewer subjects decide to work in Treatment 2. However, this difference is not significant in a two-tailed test (z=-1.48; p= 0.132), while in our regression analysis we find a significant difference on a 10% level.

Table 2. *Work Decisions Comparing Treatment(s) with Control*

Dependent Var Work	**Social-Image Game** Logistic Reg OR (SE)	**Self-Image Game** Logistic Reg OR (SE)
Treatment 1	5.393*** (2.045)	—
Treatment 2	3.467*** (1.341)	—
Treatment (Self-Image without Observer)	—	3.747*** (1.601)
Constant	0.1235*** (0.039)	0.075*** (0.029)
Observations	300	300
Pseudo R^2	0.071	0.045
LR χ^2=(df)	24.57 (2)	12.04 (1)

Notes: Dependent variable: Work (= 1 if the Allocator chose to complete the effort task, 0 otherwise). Reported coefficients are odds ratios from logistic regressions; robust standard errors in parentheses. The omitted reference category is Control. In the Social-Image Game, Treatment 1 is the condition where the Observer is uninformed about the voluntary nature of work, and Treatment 2 is the condition where the Observer is informed. The Self-Image Game has no Observer and therefore only one treatment condition. For the Social-Image Game: N = 300 (100 Control + 100 Treatment 1 + 100 Treatment 2). For the Self-Image Game: N = 300 (100 Control + 200 Treatment). *, **, *** indicate significance at the 10%, 5%, and 1% levels, respectively.

As a robustness check, we asked Allocators who worked whether they would still have chosen to complete the work task if they'd had to make their transfer decision *before* deciding whether to work. Few Allocators reported that they would have chosen to work if the work task came after the transfer decision: 5/44 in the Self-Image Game, 6/40 in the Social-Image Game in Treatment 1, and 3/22 and 6/30 in Treatment 2. This result further supports the findings from the *Control* condition—that a significantly larger number of Allocators chose to work in order to justify transferring less, rather than potentially acting on demand effects or alternative motivations.

B. Gaining Benefits: Working Allocators Reduce Transfer

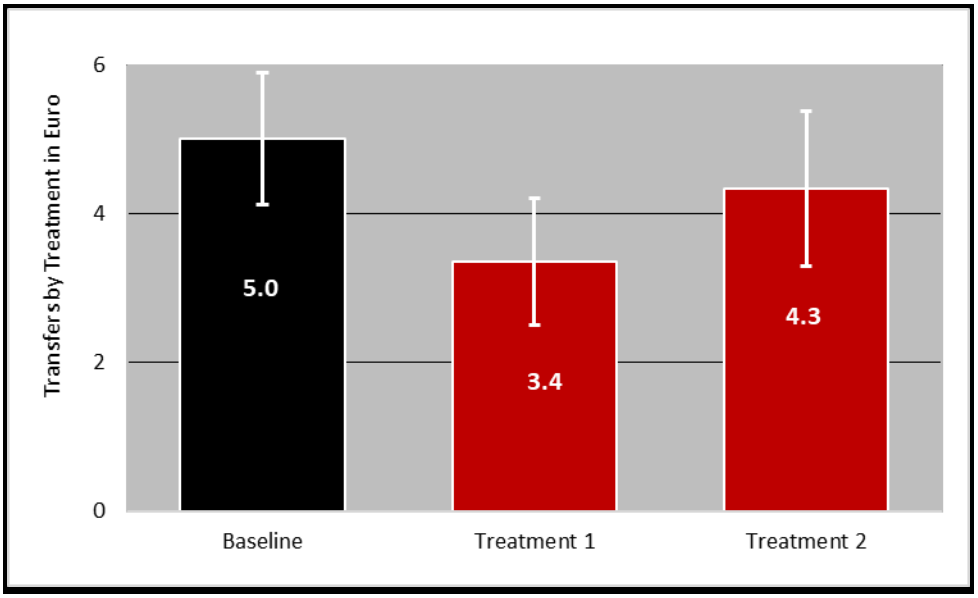
Having shown that Allocators strategically alter their decision environment, we now examine whether they exploit the resulting shift in the applicable fairness norm to obtain the self-interested benefit of reducing their transfers.

Hypothesis B_1: ***Allocators in Treatments 1 & 2, which include a work option, will transfer significantly less than those in the Baseline condition, where no work option is available.***

We tested **Hypothesis B_1** by comparing transfer amounts in Treatments 1 and 2, where Allocators had a work option, with transfers in the Baseline condition, where they did not. The data confirm B_1. In the Social-Image Game, Allocators in the Baseline (no-work) condition transfer an average of €5.00 whereas average transfers drop to €3.43 in Treatment 1 and €4.29 in Treatment 2. Mann–Whitney U tests show that transfers are significantly lower than the Baseline in both Treatment 1 (N = 100, z = 3.54, p < 0.001) and Treatment 2 (N = 100, z = 2.25, p = 0.021).[11] The graph below depicts the average transfers across treatments.

[11] We also observe that transfers are significantly higher in Treatment 2 than in Treatment 1 (z = -1.906, p = 0.05). We will further analyze and interpret this result in a separate section.

Figure 1. *Average Transfers of Baseline, Treatment 1 and 2* in the *Social-Image Game*

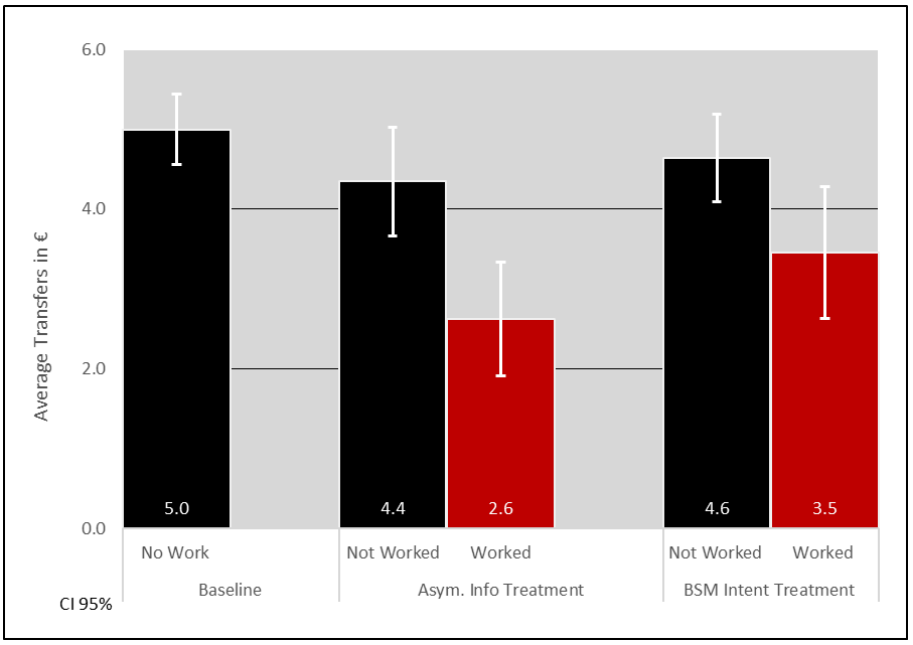


In the Self-Image Game, Allocators in the Treatment condition (N=200) transfer € 2.96 on average, significantly less than the € 4.50 transferred in the Baseline (Mann–Whitney U: $z = 5.179$, $p < 0.001$), confirming $\mathbf{B_1}$ also for a self-image motivation**.**

Next, we directly compare the transfers of Allocators who chose to work with those who did not. We posit:

Hypothesis B_2: ***Across treatments, Allocators who work will transfer significantly less than those who did not.***

Figure 2. *Average Transfers by Treatment and Work in the Social-Image Game*



In both treatments of the Social-Image Game, Allocators who worked transferred significantly less than Allocators who did not work. In Treatment 1, working Allocators transferred an average of €2.50, whereas non-working Allocators transferred €4.40 (N = 100; z = 3.72; p < 0.001). The transfers of Allocators who worked thus were only 56% as high as those who did not work. In Treatment 2, the corresponding averages are €3.50 and €4.60 (N = 100; z = 5.71; p < 0.001). Thus, working reduced transfers by 24%. Figure 2 above shows the average transfers by work.

We confirm the effect of working on transfers with a Tobit model that accounts for the censoring of transfers at the lower bound of €0 and the upper bound of €12 (the full endowment). In the Social-Image Game, we set Treatment 2 as the reference category so that we can test whether transfers in Treatment 1 and Treatment 2 each fall below Baseline, and whether Treatment 1 transfers fall below those in Treatment 2. Under this coding, the Baseline indicator is β = 0.745 € (p < 0.05), meaning Baseline transfers exceed Treatment 2 transfers by an estimated 0.745 €. The Treatment 1 indicator is β = –0.926 € (p = 0.01), indicating that Treatment 1 transfers are an estimated 0.926 € lower than those in Treatment 2. Consequently, moving from Baseline to Treatment 1 entails a total drop of β 0.745 + β 0.926 = β 1.671, confirming that transfers in Treatment 1 are significantly lower than in Baseline. Full regression results are reported in **Table 3** below.

Table 3. *Transfers in Baseline and Treatments*

Dependent Var	**Social-Image Game**	**Self-Image Game**
Transfer	Tobit Reg (β, SE)	Tobit Reg (β, SE)
Baseline	0.745**	1.871***
	(0.371)	(0.328)
Treatment 1	-0.926**	—
	(0.373)	
Constant	4.201**	4.403***
	(0.263)	(0.266)
Pseudo R^2	0.014	0.023
LR χ^2(df) =	19.52 (2)	30.99 (1)
Censored left	38	50
Censored right	1	0
Observations	300	300

Notes: Dependent variable: Transfer (€0–12 sent by the Allocator to the Recipient). Coefficients are from Tobit regressions with censoring at $0 (lower bound) and €12 (upper bound). Standard errors in parentheses. Treatment 2 is the reference category. Baseline is a dummy equal to 1 for the Baseline condition (no work option) and 0 for Treatment 2. Treatment 1 is a dummy equal to 1 for Treatment 1 and 0 for Treatment 2 in the Social-Image Game; it is omitted from the Self-Image Game column, which has a single treatment condition. "Censored left/right" report the number of observations at the lower and upper bounds. N = 300 (100 Baseline + 200 Treatment(s)). *, **, *** indicate significance at the 10%, 5%, and 1% levels, respectively.

As a robustness check, we asked working Allocators in the Social-Image Game how much they would have transferred had they not worked. In Treatment 1, 38 of 40 Allocators, and in Treatment 2, 29 of 30 Allocators reported a higher hypothetical transfer in the absence of the effort task. Using these self-reported counterfactual transfer amounts, we recalculated average transfers to approximate what Allocators in both Treatments would have transferred if the work option was not available. Replacing actual transfers of working Allocators with the hypothetical "no-work" amounts they indicated raises treatment means to €4.7 in Treatment 1 (z = 0.291, p = 0.770) and €5.4 in Treatment

2 (z = 0.492, p = 0.624). In both cases of the hypothetical means, the difference from the Baseline average transfer (€5.03) is statistically insignificant according to Mann–Whitney tests. In the Self-Image Game the replacement yields €4.21 with no significant difference to Baseline (z = 0.271, p = 0.324).

These findings suggest that, had subjects not worked, their transfers would have closely resembled the Baseline condition—supporting the interpretation that effort causally lowers transfers by rendering applicable the more favorable effort-based fairness norm. Because these counterfactuals are elicited within-subject, they hold individual heterogeneity constant, providing further support for our claim that it is the act of working—not selection—that reduces transfers.

C. Protect Self-Image and Social-Image: Comply with Norm-Expectation

We have seen support for the first two theory elements: **(A)** that subjects seek to change their decision environment through their work, and **(B)** that they do so for personal gain by reducing their transfers. Now we turn to the third element: **(C)** that Allocators change the decision environment through their work in a way that allows them to protect their self-image and social-image.

1. Self-Image Motivation

In the Self-Image Game, Allocators' work decision cannot alter others' views of their transfer choices: There is no Observer and the Recipient is not informed that the Allocator worked. Thus, they can only be motivated by protecting their self-image in deciding to work, but not their social-image (Faillio, Rizzoli & Tontrup 2019; Hoffman et al. 1996). If Allocators think that work enables them to change the decision-making environment so that the appropriate norm is the effort-norm, it would allow them to retain more, compared to the equality norm.

To test this, we elicited whether Allocators expected work to lower the transfer amount a hypothetical third-party would regard as a fair transfer **(a)**. We then tested whether Allocators who work comply with their expectation about the effort-based norm they expect to have created through their work **(b)**. Specifically, we test whether Allocators seek to protect their self-image by matching or exceeding the transfer amount they believe Observers consider fair.

a. Hypothetical Norm Expectations

Hypothesis C_1: ***Allocators expect that work will lower the transfer amount that hypothetical others will consider as fair.***

Our results support this hypothesis. We compare, **within subjects**, the fairness norm expectations Allocators indicate conditional on working with the expectations they report conditional on not working. The former (€2.60) is significantly lower than the latter (€4.70), as confirmed by a Wilcoxon signed-rank test (z= –8.194, p< 0.01). We can further support C_1 with a **between-subject** comparison: In the Baseline condition, where no work option is available, norm expectations are significantly higher (€4.50) than in the treatment condition (€2.60), when Allocators assume they have worked (Mann–Whitney z = –5.244, p< 0.01).

b. Compliance with Norm Expectations

We now test whether working Allocators protect their self-image by complying with their expectations of the new effort-based norm.[12]

Hypothesis C_2: ***Allocators will adjust their transfers to align with what they believe third parties consider fair conditional on work.***

Allocators who work comply if their transfers fall within the range that they believe—with full (100%) or partial (50%) confidence—**hypothetical** third parties would consider a fair transfer contingent on work. Conversely, if Allocators make a transfer below this range, they subjectively violate their own expectation of what they imagine third parties would believe the new norm is and thus would not succeed in using work to preserve their self-image costs while making low transfers.

To analyze whether Allocators follow their own norm expectation, we construct a variable that considers only the norm expectation that Allocators hold with respect to the actual choice they make: that is, for those Allocators who worked, we use their norm expectation for the work scenario; for those who did not work, we use their expectation for the no-work scenario. To find out whether they followed their norm expectation, we subtract from this amount each Allocator's actual transfer. If the outcome is negative, it indicates that an Allocator`s transfer fell short of their own norm-expectation: i.e., the Allocator transferred less than what he or she expects third parties consider a fair transfer. If the outcome is zero or positive the Allocator met or exceeded their own norm expectation.

We find that 77.3% of the Allocators who decided to work in the Self-Image Game made transfers that met or exceeded their own norm expectation. If working did not systematically affect compliance with one's own norm, workers should be equally likely to choose transfers that meet (or exceed) their norm as transfers that fall short (p = 0.5). In contrast, 17 of 22 workers (77.3%) complied with their norm, a share significantly above 50% (z = 2.56, p = 0.011), indicating that compliance is not consistent with random behavior.[13]

We conclude that self-image concerns motivate both their decision to engage in BSM and the amount they ultimately transfer.

2. Social-Image Motivation and Objective Norm Shift

Norm-shifting through effort is a particularly effective BSM strategy because it not only can ameliorate Allocators' self-image costs but also their social-image costs: Allocators can preserve both their expected and actual social-image if they, and others, believe that completing the work task shifted the fairness norm from equal distribution to one based on effort, and then adhere to that shared norm they expect to have created.

Accordingly, we first analyze whether Allocators believe that their work will lower the transfer amount Observers consider to be fair **(a)**. Then, we examine whether Allocators who work comply with what they believe the effort-based norm is in the eyes of Observers **(b)**. Third, we analyze

[12] Since we assume Allocators to act behaviorally rational, we also expect them to place their transfers within the range of transfer amounts they believe is consistent with the effort-based social norm (i.e., the range of transfers they expect others to conclude would be fair).

[13] Allocators who worked also complied with their norm-expectation at a significantly higher rate than non-working Allocators (77.3% versus 50.0%; two-proportion z-test z= 3.53 p< 0.001).

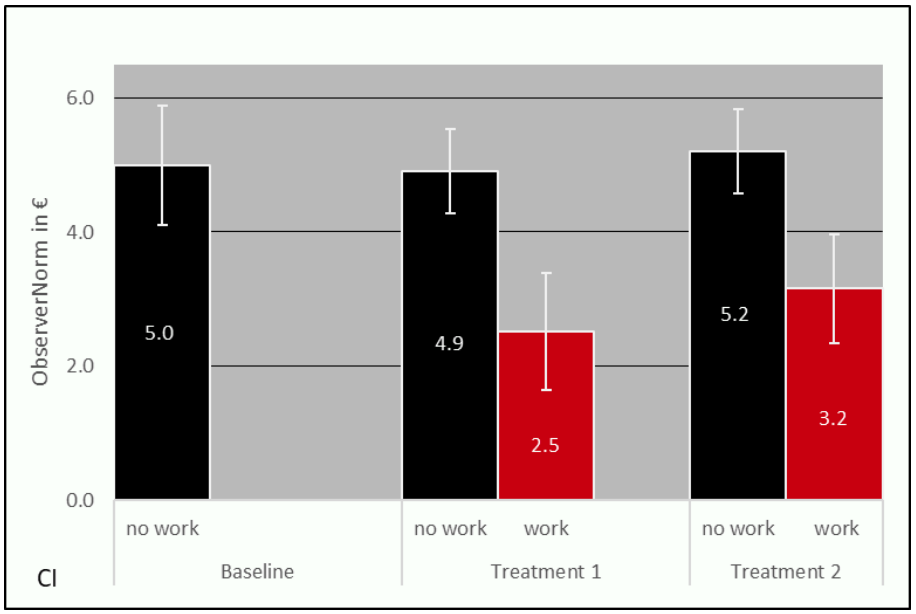
whether Observers also believe that the norm has shifted and whether Allocators' and Observers' norm expectations align **(c)**. Finally, we analyze whether Allocators' actual transfers match Observers' norm expectations so that they can successfully protect their social-image as they intended **(d)**.

a. Allocators' Lower-Bound Norm Expectations

We first analyze Allocators' lower-bound norm-expectations regarding what Observers may consider a fair transfer, conditioned on whether Allocators work or not. As in the Self-Image Game we expect Allocators to believe that work will lower the amount that impartial Observers expect them to transfer.

Hypothesis C_3: ***Allocators expect Observers to consider a lower transfer as fair conditional on Allocators' work.***

Figure 3. *Allocators' Lower-Bound Norm Expectation by Work and Treatment*



In both treatments, Allocators expect Observers to believe that they can fairly transfer substantially less if they work than if they do not: in Treatment 1 €2.6 as compared to €4.9 (Wilcoxon signed-rank test z= –5.654, $p< 0.01$) and in Treatment 2 €3.2 as compared to €5.2 (Wilcoxon signed-rank test (z= –5.934, $p< 0.01$). We can further support C_3 by comparing subjects' norm expectations in the Baseline condition, where no work option is available, with Allocators` norm expectations in both Treatments: In Baseline norm-expectations are significantly higher (€5.0) than in Treatment 1 (€2.50; Mann–Whitney $z= –4.265$, $p< 0.01$) and Treatment 2 (€3.2; Mann–Whitney $z = –3.736$, $p < 0.01$), where Allocators condition their norm-expectation on work.

For illustration *Figure 3* above shows Allocators' average norm expectations by treatment and work decision.

b. Allocators Comply with their own Norm Expectations

As in the Self-Image Game, our BSM hypothesis suggests that Allocators who work should align their transfers with their expectations about what Observers consider a fair transfer for a working Allocator.

Hypothesis C_4: ***Allocators who work adjust their transfers to their norm expectations.***

Allocators comply if their transfers fall within the range that they expect—with either full (100%) or partial (50%) confidence—Observers will consider fair. Conversely, Allocators who make a transfer below this range subjectively violate their norm expectation.

As in the Self-Image Game, we construct a variable that measures the difference between Allocators' actual transfers and their norm expectations conditional on whether they worked. Thus, for Allocators who worked, we subtracted their norm expectation given work from their actual transfer; for those who did not work, we subtracted their norm expectation from transfers given no work. A negative value indicates that the Allocator`s transfer fell short and violated their norm expectation; if the outcome is zero or positive the Allocator met or exceeded their norm expectation.

If working did not systematically affect compliance with one's own norm expectations, workers would be no more likely to choose transfers that meet (or exceed) their norm than transfers that fall short. We therefore use $p = 0.5$ as a statistical benchmark representing a situation with no tendency in either direction and test whether the observed compliance rate among workers exceeds this value. In Treatment 1, 35 out of 40 Allocators who worked complied with their norm expectations (87.5%), which is significantly above $p= 0.5$ (one-sample proportion test: $z= 4.74$, $p < 0.001$). Similarly, in Treatment 2, 28 out of 30 workers (93.3%) complied with their norm expectations, again far above the 50% benchmark ($z= 4.75$, $p < 0.001$).[14]

c. Observers' Judgment of the Norm

We now turn to analyze the Observers' norm perception to support our hypothesis that by accepting the work task, Allocators not only adjust their own norm expectation, but also shift what impartial Observers consider a fair transfer.

Hypothesis C_5: ***Observers should believe the fairness norm to be less demanding for Allocators who completed the work task.***

For this analysis, we asked Observers (N=35 for each treatment) what they believed the fairness norm should be when Allocators had worked and when they had not worked. For each possible transfer, Observers indicated whether they thought it met or fell short of the fairness norm. We asked Observers to state their confidence that the transfer is fair, selecting whether they were 100% confident, 50% confident, or had no confidence, i.e. thought the transfer amount was unfair. Second, Observers assigned all, half, or none of a 50-cents reward to all possible transfers of Allocators depending on how confident they were that the transfer was fair. For a transfer they were fully confident met the fairness norm, they were instructed to assign the full 50-cents, a transfer they were 50% confident sure met the norm, was to receive 25-cents**.** An Allocator making a transfer they thought was unfair was to receive nothing.

[14] In Treatment 1, 87.5% of Allocators who worked complied with their norm expectations, compared to 73.7% of those who did not work ($\chi^2 = 3.03$, $p = 0.082$). In Treatment 2, 93.3% of workers complied with their norm expectations, compared to 67.1% of non-workers ($\chi^2 = 7.18$, $p = 0.007$).

Figure 4. *Range of Observer-Norm Conditioned on Work & Treatment*

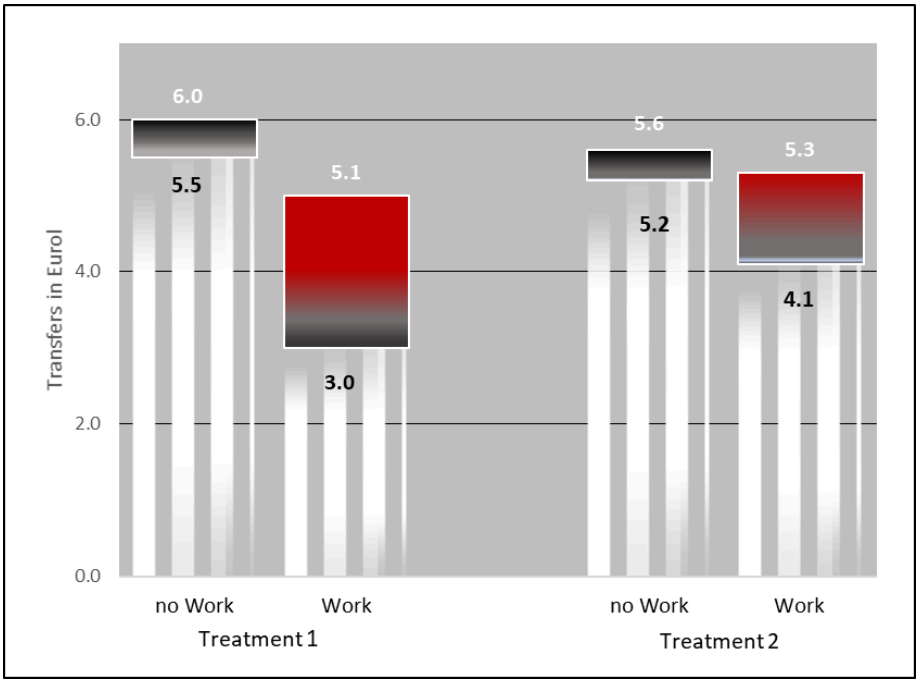


Figure 4 above illustrates the range of the fairness norm from the Observers' perspective. The lower bound represents the minimum transfer amount that Observers indicated they had 50% confidence that it would meet the fairness norm; the upper bound represents the minimum transfer that Observers were 100% confident would meet the fairness norm. The graph presents these norm boundaries averaged across Observers, separately for cases where Allocators worked versus did not work.

In support of $\mathbf{C_5}$ we find that Observers set the fairness norm significantly lower conditioned on that Allocators completed the work task compared to assuming they had not, both in Treatment 1 (signed rank test $z = -5.884$ $p < 0.01$) and in Treatment 2 (signed rank test $z = -2.720$; $p < 0.01$).

d. BSM Preserves Social-Image: Transfers Match the Norm

Work as a BSM strategy can reduce social-image costs only if Allocators' actual transfers are consistent with what Observers actually consider a fair transfer. For brevity in reporting results, we pool data from Treatments 1 and 2.

Evidence of Allocators meeting the true Observers' norm expectation would reveal that Allocators' norm-shifting strategy is effective in diminishing not only perceived but also **actual** social-image costs. This is important because norm-shifting would not be a dynamically robust strategy if Allocators who worked made transfers they expected Observers to consider fair and then later discovered that most Observers who they encountered actually believed they made unfair decisions, warranting social disapproval.

To analyze this question, we use the same style of analysis as for subjective norm compliance but here subtract the Allocators' actual transfers from Observers' (rather than their own) norm

expectations. A negative result indicates that Observers would consider the Allocators' transfers unfair, whereas a zero or positive result indicates that the Allocators met or exceeded the norm. To assess whether norm adherence is significant, we conduct a binomial test against the assumption that both the Observers' assessment and the Allocators' transfers are randomly distributed.[15]

Hypothesis C_6: ***The transfers of Allocators who work will match or exceed the minimum amount Observers judge to be potentially fair significantly more often than would be expected by chance in our Allocator–Observer pairs.***

Our data show that 68.1% of workers' transfers met or exceeded the average of Observers' minimal norm-expectations. The binomial test indicates that the observed rate of objective compliance (68.1%) is significantly higher than the random benchmark of 0.5 ($p < 0.01$).[16]

The social-image costs Allocators would actually experience outside the lab should depend on the percentage of third-parties they encounter who conclude that the Allocator made a fair or an unfair transfer. In order to test whether workers' actual transfers line up with the effort-based fairness norm of Observers, we compared the transfer of every Allocator who worked, with each Observer's norm expectation, specifically, the minimum transfer that the Observer rewarded with 25 Euro cents or more. For each Allocator, we then determine the percentage of Observers whose fairness norm was satisfied by the Allocator's transfer. We then average these percentages across Allocators. With this Cartesian comparison, we can show how widely Observers accept the transfers that Allocators made. The forty workers in Treatment 1 satisfy on average 64.3% of the Observer sample, while the thirty workers in Treatment 2 satisfy 63.0%. Pooling the sixty-nine workers across both treatments yields an overall alignment of transfers and expectations of 63.7%. We conclude that actual transfers meet nearly the lower-bound norm expectations of two-thirds of Observers.

Importantly, Allocators and Observers broadly share the judgment that completing the work shifts the relevant fairness norm: more than 70% of participants in both roles report complete confidence that the work effort changes what is considered fair. This indicates that working establishes a new *effort-based* fairness norm. As a result, Allocators who work can keep more than an equal share without being judged unfair by most Observers, substantially reducing the risk of social disapproval.

While participants played the game only once, we have reason to expect that the results are dynamically stable: Allocators demonstrate a strong BSM motivation and comply at very high rates with what they expect are Observers' true norm expectations, suggesting that they would adjust to feedback in repeated interaction. That is, if they learned that their transfer fell short of Observers' actual norm expectations, they might increase transfers to protect both their self-image and social-image. The dynamic stability of BSM strategies presents an interesting question for future research.

[15] Alternatively, we could use a rational-choice benchmark with Allocators transferring nothing, setting a probability of 0.
[16] For non-workers we observe an objective compliance rate of 55.7%.

D. Strategic Behavior: Allocators' Work & Transfer Choices are Behaviorally & Strategically Rational

The fourth element of our BSM theory states that individuals strategically engage in norm-shifting to reduce their self- and social-image costs when retaining more than half of the allocation **(D)**. Thus, the greater the self-image and social-image costs Allocators can avoid by engaging in the work task, the more likely they should be to work.

We assume that self-image and social-image costs from making transfers that violate the applicable fairness norm are heterogeneous and increase with subjects' prosociality.[17] Proself types can make low transfers without working because they incur low self- and social-image costs when they violate a fairness norm. By contrast, prosocials face high self- and social-image costs from norm violations and thus would make higher transfers absent work. Accordingly, they stand to gain more from using BSM to shift the norm to one under which they can fairly retain more than half of the allocation.

Allocators' strategic choice of whether to engage in BSM should also depend on the extent to which they expect completing the effort task to lower the fairness norm Thus, Allocators who are equally prosocial may still face very different incentives to work depending on their norm expectations: they will choose to work if they expect that working shifts the fairness norm sufficiently to lower their transfer such that the resulting monetary gain outweighs the combined effort and image costs of using BSM. By contrast, if they expect that the resulting norm shift will be too small to offset BSM costs, even prosocial types have no incentive to work.

1. Work Choices

Thus, our assumption that Allocators make strategic, behaviorally-rational work decisions leads to two hypotheses set forth below.

Hypothesis D_1: ***The more prosocial an Allocator is, the more likely they are to choose to work.***

Hypothesis D_2: ***The lower Allocators expect others' beliefs about the effort-based norm to be, the more likely they are to work.***

As these two hypotheses are interrelated, we estimate a logistic regression with "Work*"* as the dependent variable and the key predictors of "Prosociality" and Allocators' "Norm Expectations". We find significant support for both hypotheses.

In the Social-Image Game we find that prosociality (D_1: OR = 1.015; p = 0.02) and norm-expectations (D_2: OR = 0.568; $p < 0.01$) significantly increase the odds that Allocators take up the work. This means that, holding Allocators' norm expectations constant, the more prosocial an Allocator is the more likely she is to work. In turn, holding prosociality constant the higher Allocators' norm expectations are, the less likely they are to work.

[17] Prior studies involving social preferences have found that transfers are highly correlated with individuals' prosociality: e.g., dictator games (Murphy et al., 2011; van Lange, 1999); trust games (Bochet et al., 2006; van den Bos et al., 2009; Fetchenhauer & Dunning ,2012); and public goods games (Balliet et al., 2009; Bogaert et al., 2008; Fiedler et al., 2013).

Table 4. *Work Decisions by Norm Expectations and Prosociality*

Dependent Var Work	**Social-Image Game** Logistic Reg OR (SE)	**Self-Image Game** Logistic Reg OR (SE)
SVO	1.015**	1.035***
	(0.007)	(0.008)
Norm-Expectation	0.568***	0.499***
	(0.081)	(0.084)
Treatment 2	0.549*	—
	(0.180)	
Constant	2.341*	0.629
	(1.187)	(0.311)
Observations	200	200
Pseudo R^2	0.096	0.189
LR χ^2(df)=	23.60 (3)	40.48 (2)

Notes: Dependent variable: Work (= 1 if the Allocator completed the effort task). Reported coefficients are odds ratios from logistic regressions; robust standard errors are in parentheses. SVO denotes the Allocator's Social Value Orientation, measured as a continuous angle in degrees (see Appendix). Norm-Expectation is the minimum transfer the Allocator expects Observers to consider fair (at 50% confidence). In the Social-Image Game column, Treatment 2 equals 1 (strategic intent revealed) and 0 for Treatment 1 (reference category); Treatment 2 is omitted from the Self-Image Game column, since for this Game we have only a single Treatment with the design of Treatment 2. *, **, *** denote p < 0.10, 0.05, and 0.01, respectively.

In the Self-Image Game, we find the same pattern of results: prosociality (D_1: *OR*= 1.035; $p<0.01$) increases the odds of working, and Allocators' norm expectations (D_2: *OR*= 0.499; $p<0.01$) significantly decreases the odds of work decisions.

Above we reported logistic regression results based on odds ratios. To make the magnitude of the effects more intuitive, we also estimate linear probability models (LPMs) with Work as a 0–1 dependent variable. Because the outcome is binary, the LPM coefficients can be read as approximate changes in the probability of working in percentage points for a one-unit change in the regressor.

In the Social-Image Game, the coefficient on SVO (Angle) is 0.0053. This means that each additional 1° of prosociality increases the probability of working by about 0.53 percentage points. A moderately prosocial Allocator with SVO = 45° is therefore roughly 24 percentage points more likely to work than a purely self-interested Allocator with SVO = 0° (0.0053 × 45 =0.24).

The coefficient on Norm-Expectation (the minimum transfer the Allocator expects Observers to deem fair when she works) is −0.098. Thus, each additional €1 in the expected minimum fair transfer reduces the probability of working by about 9.8 percentage points. If an Allocator believes that working relaxes the minimum fair transfer from €6 to €3, the probability of working rises by about 29–30 percentage points (3 × 0.098 = 0.29).

In the Self-Image Game, the SVO coefficient is 0.0028, implying that each extra degree of prosociality raises the probability of working by about 0.28 percentage points. Moving from SVO = 0° to SVO = 45° therefore increases the probability of working by roughly 13 percentage points. The Norm-Expectation coefficient is −0.108, so each €1 increase in the expected minimum fair transfer lowers the probability of working by about 10.8 percentage points. An Allocator who believes that working reduces the minimum fair transfer from €6 to €3 is therefore about 32 percentage points more likely to work than someone who believes the minimum remains at €6.

When we add up these effects, in the Social-Image Game a moderately prosocial Allocator (45°) who expects a large norm shift from €6 to €3 is predicted to be roughly 24+29 = 53 percentage points more likely to work than a purely self-interested Allocator (0°) who expects no shift. In the Self-Image Game, the corresponding difference is about 13+32 = 45 percentage points. We use the LPM only for intuition, but these large probability gaps line up closely with the logit estimates and reinforce the conclusion that both prosociality and, especially, the expected norm shift strongly increase the likelihood that Allocators engage in BSM by choosing to work.[18]

2. Transfer Decisions

As noted above, the literature shows that the more prosocial an individual is, the more they tend to transfer to the Recipient and the more likely they are to comply with fairness norms; by contrast, proself subjects may simply breach the norm without incurring strong self-image costs. This gives prosocials not only stronger incentives to work, but also greater scope than proself subjects to reduce their transfers by undertaking work because they are committed to higher transfers absent work. However, while we expect prosocial Allocators to reduce their transfers more when they work, they may still allocate more to the Recipient than proself types do.

Hypothesis D_3: ***The more prosocial an Allocator is, the more the work choice should reduce their transfer.***

As we have seen in the context of work choices, prosociality and norm expectations are interrelated. The extent to which a prosocial individual is willing to reduce transfers depends on what they expect the Observer to regard as fair, conditional on work completed. This leads to our second hypothesis:

Hypothesis D_4: ***The lower the transfer amount an Allocator expects the Observer will consider fair conditional on work, the less that Allocator will transfer.***

As reported in the literature, prosocials generally transfer more than proselfs when no work option is available, as in our Baseline. Our results for the Social-Image Game, €4.13 and €6.2 (Mann-Whitney z= -4.912; p <0.01), and for the Self-Image Game, €3.50 and €5.90 (z =-6.034; p<0.01) are consistent with these previous findings.

In support of Hypothesis $\mathbf{D_3}$, we find that working prosocial Allocators reduce their transfers more than working proselfs. Notice that the common binary classification of prosociality that we use, splits subjects into prosocial and proself types at 22.5° on the ring measure (see Appendix). An SVO of 22.5° however, still reflects meaningful other-regarding concern: these individuals are willing to increase the other's payoff at some cost to themselves, just less strongly than prosocial types with higher angles. This is why we also expect proself subjects to choose work and to reduce their transfers when they work—just less often, and by a smaller amount, than prosocials.[19]

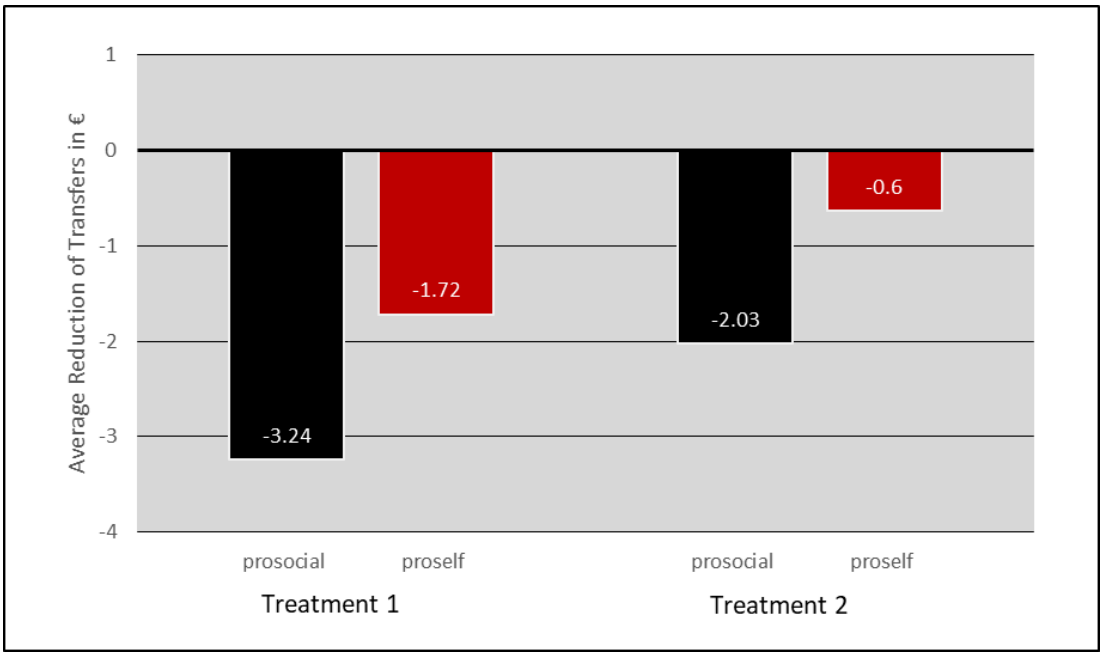
[18] To interpret our regression results directly as probabilities we could use a linear probability model. However, the model is not as accurate as a logistic regression for binary outcomes (like the choice whether or not to work). The regression yields a significant difference of 15.4%. (β= 0.154** SE=0.053).

[19] By contrast, subjects at 0° show no other-regarding concern, and subjects with SVO below 0° are competitive, willing to sacrifice their own payoff in order to stay ahead of others.

In Treatment 1 of the Social-Image Game, the reduction is €1.91 for prosocials and €1.27 for proselfs (z= –4.912, p < 0.01). In Treatment 2, the reduction is €1.35 for prosocials and €0.38 for proselfs (z= 5.871, p < 0.01).

In the Self-Image Game we find a difference in transfers of €2.76 for prosocials and €1.76 for proselfs (z= –6.112, p< 0.01). **Figure 6** shows the descriptive results of the Social-Image Game.

Figure 6. *Work Reduces Transfers by SVO-Type and Treatment*



To confirm these results, we perform a logistic regression (see **Table 5** below). The interaction term (SVO*x*Work) confirms that the negative impact of work on transfers is larger for prosocials than for proselfs, while the SVO variable indicates that prosocials generally give more than proselfs as expected.

Table 5. *Prosociality and Norm Expectations Impact Transfers*

Dependent Var Transfer	**Social-Image Game** Tobit Reg (β, SE)	**Self-Image Game** Tobit Reg (β, SE)
Work	-0.180	-0.715
	(0.549)	(0.711)
SVO	0.616***	0.068***
	(0.008)	(0.009)
SVO*x*Work	-0.049***	-0.050***
	(0.015)	(0.0167)
Norm Expectation	-0.559**	-0.371***
	(0.126)	(0.127)
Treatment 2	1.157***	—
	(0.323)	
Constant	0.614	1.587***
	(0.539)	(0.454)
Censored left	29	48
Censored right	0	0
Observations	200	200
Pseudo R^2	0.104	0.101
LR χ^2=(df)	93.92 (5)	87.22 (4)

Notes: Dependent variable: Transfer (€0–12 sent by the Allocator to the Recipient). Coefficients are from Tobit regressions with censoring at \$0 (lower bound) and €12 (upper bound); standard errors (in parentheses). The sample includes only treatment observations (Control is excluded because it has no work option). Work is an indicator equal to 1 if the Allocator completed the effort task and 0 otherwise. SVO denotes Social Value Orientation (continuous angle in degrees; see Appendix). SVO×Work is the interaction term; its negative coefficient indicates that, among workers, more prosocial Allocators reduce transfers more than less prosocial workers. Norm Expectation is the Allocator's belief about the minimum transfer that Observers consider fair, conditional on work. In the Social-Image Game column, Treatment 2 equals 1 (strategic intent revealed) and 0 for Treatment 1 (reference category); Treatment 2 is omitted from the Self-Image Game column, since for this Game we have only a single Treatment with the design of Treatment 2.

E. Social-Image Costs of Using BSM

Finally, we ask whether engaging in BSM itself may entail a social-image cost, when it is transparent that Allocators chose to work solely in order to shift the norm for a self-interested gain and to the detriment of the Recipient. Given its self-interested intent, we expect norm-shifting in our study to cause some social-image costs with Observers, even though they agree that an effort-based norm applies when Allocators complete the work.

This notion of social "BSM costs" (our label) finds support in work by Bartling et al. (2014) who implement a binary dictator game in which the Recipient's outcomes are initially concealed. Building on Dana et al. (2007), Allocators could uncover Recipients` outcomes before selecting an allocation or remain ignorant. Bartling et al. give third parties the option to punish Allocators for blindly choosing an allocation that is in their favor. Observers apparently interpret the blind decision as an attempt to protect one's self-image while acting in self-interest and punish Allocators.

We test for evidence of BSM costs by comparing our two Treatments in the Social-Image Game. In Treatment 2, Observers know Allocators had a choice whether to work, making the self-interested BSM motivation salient. In Treatment 1, Observers are not informed that working was voluntary, concealing the Allocators' self-interested motivation for completing the task. In a post-experimental manipulation check we confirm that as intended by our design, a large majority of Observers in

Treatment 1 indicates that they assumed Allocators were instructed to work rather than having chosen to work to benefit from shifting the norm in their favor.

Hypothesis E_1: ***Observers' norm expectations, contingent on work, should be higher in Treatment 2 than in Treatment 1.***

We find support for our hypothesis. While Observers in both treatments believe that Allocators who work can fairly retain more than those who do not, we see a difference between the treatments: Whereas Observers in Treatment 1 indicate that Allocators who work should transfer a minimum of €3.03, Observers in Treatment 2 belief that working Allocators should transfer at least €4.14. The difference is significant (Mann Whitney U z =2.235; p =0.015). Thus, individuals who openly engage in BSM to reduce the normative demands placed on them face social-image costs in the form of a less generous effort-based norm.

We expect Allocators to anticipate that Observers in Treatment 2 will take into account their self-interested intent and accordingly set a higher fairness norm (see Cushman 2008; Falk et al. 2003). As Allocators who work demonstrate that they are motivated to comply with what they believe are Observers' norm expectations, we posit:

Hypothesis E_2 : *Allocators who work should transfer more in Treatment 2 than Treatment 1.*

Our results support this hypothesis. In Treatment 1 we find workers transfer an average of only €2.50, whereas in Treatment 2 they transfer with €3.41 significantly more (z = -2.478; p =0.01). When we look at the full sample, we find an average transfer of €3.43 in Treatment 1, whereas transfers in Treatment 2 are significantly higher, €4.29 (z =-1.906; p =0.04).

Figure 2 above illustrates that the transfer difference between the treatments is driven by workers.

We confirm our results using a Tobit regression that allows us to control for Allocators' prosociality and their norm expectations. We estimate the interaction term of Treatment 2 × Work and find it significant at the 1% level, showing that working Allocators transfer significantly more in Treatment 2 where their strategic BSM intent is salient (see **Table 6** below). Including the interaction term of SVO×Work does not meaningfully affect results. For parsimony, we omit this interaction here

We conclude that individuals who openly engage in BSM face social-image costs for strategically shifting the norm in their favor. Importantly, in turn, our result also suggests that the social-image cost of using BSM can diminish if the use of BSM intent is concealed. Concealing the use of BSM may often be an integral part of an effective BSM strategy. For example, an actor may conceal the BSM intent of the work choice by selecting a form of work that benefits their company or society, thereby providing a non-BSM explanation for their work choice. Or a principal may delegate a self-serving task to an agent in order to reduce own accountability and conceal the use of BSM by framing the task as an inherent part of the agent's regular work. This framing may allow the decision-maker to benefit from BSM while avoiding responsibility for employing it.

Table 6. *BSM Costs in Transfers*

Dependent Var	**Social-Image Game**
Transfer	Tobit Reg (β, SE)
Work	-2.261***
	(0.494)
Treatment 2	0.734*
	(0.404)
Treatment 2*x*Work	1.351**
	(0.677)
SVO	0.047***
	(0.007)
Norm Expectation	0.561***
	(0.128)
Constant	1.167**
	(0.563)
Censored left	29
Censored right	0
Observations	200
Pseudo R^2	0.100
LR $\chi^2(5)$=	87.22

Notes: Dependent variable: Transfer (€0–12). Coefficients are from a Tobit regression with censoring at €0 and €12. Standard errors are in parentheses. Work is equal to 1 if the Allocator completed the effort task and 0 otherwise. Treatment 2 is a dummy equal to 1 and 0 for Treatment 1 (reference category). Treatment 2×Work tests whether the effect of working on transfers differs between treatments. SVO denotes Social Value Orientation (continuous angle in degrees); Norm-Expectation is the Allocator's belief about the minimum fair transfer that Observers expect, conditional on the work decision. The sample includes only Treatment 1 and Treatment 2. *N*=200. *, **, *** indicate significance at the 10%, 5%, and 1% levels.

F. Summary of Results

We find support for all four distinct elements of our BSM theory: First, people change their decision environment through work, such that a norm applies that is more in their favor. Second, our Allocators capitalize on the norm shift they have engineered, reducing their transfers in line with that new effort-based norm.

Third, Allocators decide to work even when they are not observed. They reduce their transfers to what they expect third parties would consider a fair transfer conditional on Allocators working. Since Allocators shape their behavior with reference to hypothetical third parties, we can conclude that they aim at preserving their self-image. Allocators also protect their social-image: We see evidence that the effort-based norm is shared: More than 85% of Observers and Allocators share the understanding that completing the work task effectively shifts the norm. Although the effort-based norm is broader and less certain in its lower bound, and most Allocators take advantage of this wiggle room by placing their transfers within this ambiguous range, few working Allocators outright violate what they expect to be the effort-based norm. Also, a majority of Allocators' actual transfers meet Observers` norm expectations, showing that most of the Allocators in fact manage to maintain their social-image.

As a fourth element of our BSM theory, we show that Allocators act strategically and behaviorally rationally. Those Allocators who face higher self- and social-image costs from violating the norm—prosocials—are also more likely to work than proselfs. Moreover, those Allocators who expect that working will effectively lower the fairness norm more are also more likely to work.

Finally, our data reveal that this strategic element of BSM has a social-image cost (it may also have a self-image cost, but we did not measure it): The self-interested intention to shift the norm, which is salient only in Treatment 2 of the Social-Image Game, increases what Observers think the fairness norm is and Allocators, who correctly anticipate that increase, match their transfer to this higher norm to protect their social-image.

V. Discussion

A. Internal Validity

1. Beliefs, Cognitive Dissonance and Self-Serving Bias

We analyze Allocators' norm expectations contingent on whether they undertake the effort task and show that these expectations explain both their work and transfer decisions. We elicited Allocators' expectations only after they had made all experimental choices to avoid the risk that belief elicitation would influence their decisions.

This elicitation order, however, raises two potential concerns: Allocators might have aligned their stated expectations with their actual choices to reduce cognitive dissonance (Festinger, 1957; Cooper, 2012), or they might have distorted their expectations through motivated reasoning to justify selfish transfers. We address both concerns with two empirical checks. First, we conducted a pilot study in which Allocators reported their norm expectations without first making any incentivized decisions. Pilot responses closely match main-study expectations (difference < €0.40), suggesting that neither cognitive dissonance nor self-serving bias substantially distorted stated norm expectations. Second, we compare Allocators' norm expectations with impartial Observers' actual norm perceptions. The gap is less than €0.50, indicating at most modest self-serving bias and limited cognitive dissonance effects. Both discrepancies are small relative to the approximately €2 norm shift that Allocators generate through working.

The conclusion that Allocators reported presumptively unbiased expectations also is aligned with our theory of why Allocators work: Allocators will choose to work if they assume that work will reduce social-image and self-image costs, but these costs (and social-image costs in particular) are only lowered if Allocators genuinely expect both that Observers believe work shifts the norm, and that their actual transfer choices align with Observers' beliefs. Thus, Allocators have an incentive to accurately estimate the Observers' perception of the norm and make transfers within the zone they expect Observers to believe is fair.

2. Ruling out Alternative Motivations for Allocators' Work Choice

We hypothesize that Allocators take up the work task to shift the fairness norm in their favor. To provide clean support for this hypothesis, our experimental design aimed to rule out alternative motivations such as experimenter demand effects, social desirability concerns, curiosity, or misconceptions that working would increase payment or express gratitude to the experimenter.

We addressed these potential confounds in two ways. First, by experimental design: we made specific design choices that should have rendered the mentioned alternative motivations for working unlikely. And second, empirically, by conducting a Control group, which—when compared to the

other treatments—allows us to conclude that the observed treatment differences are driven by participants' intent to shift the norm.

a. Design Choices

Experimenter Demand Effects. Our results would be confounded by demand effects if Allocators could have been motivated to work in order to benefit the experimenter. To make this potential confound unlikely to occur, we employed a double-blind experimental design: Subjects were unaware of the experimenters' identities and logged in to the study using anonymous credentials that we also used for their payment, which increased social distance between participants and experimenters (Hoffman et al., 1996).

Intrinsic and Prosocial Motivation. We also designed our experiment to reduce the risk that Allocators might be motivated to undertake the work task because they are intrinsically motivated or thought the work would benefit others. To counteract these potential confounds, we selected a tedious, uninteresting and purposeless number-counting task that obviously provided no benefit to the experimenters, the Recipients, or anyone else. The tediousness should render unlikely the possibility that completing the task itself might stimulate subjects' intrinsic motivation and the purposelessness should prevent subjects from thinking that they should complete the effort task because it is socially productive.[20]

Curiosity. We also sought to ameliorate the possibility that Allocators might decide to work out of curiosity, instead of for strategic BSM reasons, by giving subjects the opportunity to try the task to ensure them that the work task does not offer any kind of interesting hidden surprise.

Misconceptions. We made sure in the instructions that subjects understood the experiment and specifically, that the real effort task **c**ould not elevate their endowment. We tested their understanding of this crucial point in the control questions.

b. Control Treatment

To test whether our experimental design succeeded in eliminating the potential confounds, we implemented a Control treatment. In the Control treatment, Allocators received an endowment as a lump sum that they can keep and that is not to be used in the experiment. They were then instructed that they can transfer non-monetary points to the Recipients. Prior to making that choice, they were given the choice to undertake the real effort task. Our theory predicts that Allocators will not work in the Control. By contrast, Allocators should work in Control if they are motivated to work by curiosity, demand effects, or the other alternative motivations set forth above. For example, if Allocators were curious or intrinsically motivated to complete the effort task, we would expect them to be just as motivated to work and work just as often in the Control as in our Treatments. The same holds, if they worked because they believed that working would serve experimenters research and they wanted to benefit them. The Control treatment also allows us to rule out another alternative motivation: a genuine desire to earn the endowment rather than receive it as a windfall—a motive that, unlike BSM, is not a self-interested strategic attempt to earn more through working.[21] If subjects were motivated solely by a desire to *earn* their endowment through work, they could still satisfy this motive in the Control treatment.

[20] The simplicity of the effort task also eliminates skill- or knowledge-related variance across the sample.

[21] We thank Rick Brooks for this insight.

The fact that significantly fewer subjects chose to work in the Control condition—compared to our other treatments, suggests that it is the latter, strategic motivation that drives the treatment effect.

B. External Validity

1. Lab Population and Field Evidence.

Our participant sample consists of both university students and professionals who have graduated in the last ten years from the university. While professionals and students do not appear to exhibit different behaviors in our study, our subjects received more academic training compared to a representative real-world sample, which might increase the sophistication of our subjects in self-managing their moral behavior.[22] On the other hand, our subjects are comparatively young.

In the end it is unclear whether the characteristics of our subject pool make our participants more or less prone to use BSM compared to a representative population sample. Younger individuals may face stronger social-image concerns that motivate BSM, while older individuals with more life experience may be more skilled at deploying BSM strategies to manage self- and social-image costs effectively.

Evidence that indirectly supports our claim that people should employ BSM in real-world decision-making can be found in field experiments that show wiggle room behavior to be as common in the field as among a student population such as ours, in the lab (e.g., Freddi, 2021).

2. The External Validity of Dictator Games

One might question our results by noting the relatively low stakes in our dictator games and arguing that, at higher stakes, individuals would simply pursue self-interest and thus have no need to undertake effort to engage in BSM. Yet, cross-cultural evidence (Henrich et al., 2001; 2005) shows that in poorer countries—where identical nominal amounts represent far larger real gains—prosocial giving remains strikingly stable. Some studies vary stakes in the lab and find that generosity decreases proportionally as stakes increase (Schier et al., 2016). However, meta-studies (Larney et al., 2019; Engel, 2011) show that prosocial behavior in dictator games is remarkably robust across a wide range of stake sizes, so the underlying motive to engage in BSM should persist when stakes are higher.

Indeed, it seems likely that our experiment's modest payments make the experimental test more stringent: we find that participants are willing to invest effort to shift the fairness norm even for small rewards, and larger incentives may only heighten that willingness. This reinforces our conclusion that individuals are both willing and able to engage in BSM.

3. Nature of Our Effort Task

We intentionally selected an effort task that was not intrinsically motivating, and that did not increase the size of the available allocation or provide any other social benefit, in order to be able to clearly isolate BSM as a motivation for undertaking effort. By contrast, in the real world, we expect that people seeking to use work to invoke a more favorable merit-based fairness norm generally will

[22] See however Exadaktylos et al. (2013), who compare subject responses in the dictator ad ultimatum games with student and representative samples finding no different results.

find work opportunities that are valued by themselves and others. This should make BSM more attractive, for a number of reasons.

First, those who engage in BSM can benefit twice, by the direct (intrinsic) benefit and also by the use of effort as a BSM strategy. Second, decision-makers whose effort is effective in enhancing the size of the endowment are even more viewed as being entitled to keep more of the resulting reward.[23] Third, the prosocial nature of the work conceals the self-interested BSM motives, making it ambiguous whether people worked to help others or primarily to shift the norm in their favor. This ambiguity reduces image costs associated with BSM, as we have shown by comparing Treatment 1, where strategic BSM intent is concealed, with Treatment 2 where the self-interested intent is salient.

We conclude that this difference between real-world work opportunities and our effort task should strengthen the external validity of our results.

C. Implications

The headline implication of our experiment is that BSM strategies extend to the domain of fairness: decision-makers strategically employ BSM to reduce the normative demands that ethical and social norms would otherwise impose. Importantly, we find that they do so not only in settings where their choices are private and anonymous but also when they are observed. These results demonstrate that individuals are capable and willing to use BSM to mute the constraints of fairness norms.

The policy implications of our study are significant: our finding that people engage in BSM to mute the demands of fairness norms suggests that they also are likely to exploit BSM opportunities to lessen the constraints of other social norms and the law. They can use norm-shifting to establish alternative social norms of serving the organization or fellow employees, to weaken the moral costs of non-compliance with corporate laws (Arlen & Kornhauser, 2023)—or can use mechanisms like delegation to diffuse their responsibility for norm violations (Bartling & Fischbacher, 2012; Tontrup & Sprigman, 2025). Through their representatives also organizations have the opportunity to engage in norm-shifting to enable them to achieve their corporate interests without being perceived to be acting unfairly. Accordingly, our evidence challenges claims that organizations and legal policymakers can rely primarily on social norms while reducing reliance on legal rules and formal enforcement.

Our results also have important implications for experimental research, underscoring that BSM should be considered when evaluating the policy implications of studies of both fairness norms and social norms more generally. While we focused on norm-shifting through work, people seeking to reduce the constraints that social norms place on their pursuit of self-interest can avail themselves of a host of other BSM strategies. We begin by analyzing strategies of norm-shifting and then take a brief look at other BSM strategies.

1. Other Norm-shifting and BSM strategies

Norms can also be shifted through other actions beyond undertaking effort. For example, companies can engage in norm-shifting through framing, to render salient the norms they prefer to

[23] Indeed, multiple dictator game studies find that most Allocators who earned the endowment through effort share little with the Recipient (Cherry et al., 2002; Oxoby & Spraggon, 2008).

govern decisions, such as employee compensation. For instance, performance-based rewards directed mostly at top managers—as opposed to alternative fairness norms that employees might embrace, such as the equity norm based on length of tenure, or the norm of equal pay for equal hours worked. Line employees and shareholders may have expectations about the fair relationship between average worker pay and the pay of chief executive officers and may object to CEOs earning pay that dwarfs that of the employees. Companies that wish to justify substantial CEO pay often invoke a fairness norm of "payment for excellence," highlighting performance metrics and comparing CEOs to other highly paid, high-performing executives rather than to lower-paid employees. This way companies can reframe the pay debate to focus on the accepted fairness norm of equal pay for equal work of a particular type (work of a CEO), thereby placing high pay proposals within a fairness context that is friendlier to high CEO pay.

b. Shifting social norms to undermine legal compliance

While people cannot shift legal prohibitions directly, they may aim to invoke a conflicting social norm to justify profitable but unlawful conduct. Consider corporate executives at a company that would profit from its employees violating the law, for example, by bribing officials to obtain lucrative contracts. Such conduct might also benefit the executives personally if their compensation is based on corporate sales or profits (Arlen & Kraakman, 1997). Absent BSM, these executives could not indirectly encourage bribery without incurring self- and social-image costs; nor could their employees engage in it without incurring such costs (Arlen & Kornhauser, 2023; Feldman, 2014).

Employees are subject to, and seek to comply with, multiple prosocial norms. Some arise from obligations to the company itself and fellow employees. Experimental work suggests that social norms established by the company tend to be more salient than those the law seeks to create (E.g., Bazerman & Tenbrunsel, 2011; Arlen & Kornhauser, 2023). If that is correct, then directors, executives, and line managers may be able to establish a salient local social norm under which prosociality is measured by efforts to increase corporate profit by adopting policies and procedures that reward profit-seeking activities, regardless of whether the law was violated. Employees who witness law-breaking employees being rewarded when their violations boost sales, and who see compliant employees dismissed if they fail to meet their numbers, will understand that the local social norm places profits first and does not view legal compliance as a test of being a good employee (see Arlen & Kornhauser, 2023). Creating such a norm enables employees to undertake misconduct while reducing the self- and social-image costs they would otherwise incur; indeed, they may experience a prosocial boost from their efforts to serve their firm. And managers who may engage in BSM to create such a norm, enable themselves to benefit from others' unlawful actions, while reducing their self- and social-image costs.

Another strategy is to exploit that legal rules often rest on underlying social norms that protect the same normative goods. When individuals can alter these social norms, they reduce the perceived

wrongfulness of violating the legal rule itself, lowering the moral costs of non-compliance.[24] Consider again the restaurant pet ban: A pet owner violating this rule can strategically shift the applicable social norm not to harm others by encouraging customers to pet and play with their dog, providing salient evidence that other customers consent to the pet's presence and are not concerned about safety or discomfort—though health protections remain unaddressed. Even as the law is violated, non-compliance is often tolerated by other patrons and the restaurant. Both customers and restaurants may find it in their interest to accept the norm shift if it is persuasive: customers may want to bring their own pets, and the restaurant wants to satisfy its clientele. The strategy may also exploit moral wiggle room: consent may not always be genuinely voluntary, particularly when refusal might create awkwardness or conflict. Pet owners who anticipate this reluctance to dissent can strategically leverage it to facilitate norm-shifting.

c. Other BSM strategies like sharing responsibility

Evidence suggests that people's expected self- and social-image costs from unfair decisions that benefited them depend on whether they perceive themselves to be—and expect others to perceive them to be—responsible for the decision. Recognizing this, people can and do pursue self-interest by employing an agent to make the decision on their behalf, ensuring the desired outcome through the compensation and decision-making authority provided to the agent. The use of the agent alters the decision-making environment to attenuate the actor's responsibility for the ultimate unfair choice in both their own mind and in the perception of third-parties, enabling the person employing the agent to benefit from the resulting unfair allocation without incurring the same self- and social-image costs (Tontrup & Sprigman, 2025; Hamman et al., 2010; Coffman, 20011; Bartling & Fischbacher, 2008).

Voting similarly diffuses responsibility (see, e.g., Arlen & Tontrup, 2015a) and thus should also mute the self- and social-image costs of unfair decisions that result from it (see also id.). As a result, people wanting to obtain a benefit that is only available from a norm violation—such as an illegal act—may be able to achieve their goal without experiencing the same self- or social-image costs that normally would attach to those who violate the law by using agents or group-decision-making to accomplish their goal, thereby distancing themselves from being held responsible for the norm violation. Given the ubiquity of agents (including employees) and group decision-making, this suggests that people seeking to benefit from misconduct may have ample opportunity to use BSM to mute the disciplining effect of social norms, especially in the context of misconduct occurring through corporations or other organizations.

2. Legal Policy Implications of BSM

As noted in the Introduction to this Article, a number of legal scholars have premised policy recommendations on the assumption that people can be relied on to conform to social norms **and the law**. These scholars have argued that the law should **reduce its reliance** on sanctions and enforcement, **instead relying** more on the expressive power of legal norms, which provides a less

[24] Similarly, in Europe, smokers have been observed asking others for consent to smoke when smoking is prohibited but enforcement is rudimentary (Depoorter & Tontrup, 2024). By seeking consent, smokers invoke the value of individual autonomy—the idea that people can decide for themselves whether to accept smoke exposure—thereby appearing to satisfy the social norm against imposing harm on others. Because preventing non-consensual harm is central to the ban's purpose, this behavior can also be seen as meeting the law's intent even while violating its letter. Just as in the pet example, other smokers and accommodating establishments like bars are willing to accept the norm shift if it sufficiently reduces the social pressure from other customers and the image concerns associated with violating the ban.

socially costly deterrent (see Sunstein, 1996b; Dau-Schmidt, 1990; Cooter, 1998). Our results counsel caution for those contemplating following such recommendations, at least in situations where the relevant actors can avail themselves of BSM strategies to facilitate their pursuit of self-interest.

For example, Blair and Stout (2001) assert that corporate law should not employ the threat of legal liability to deter powerful shareholders, officers, and directors from opportunistically seeking private benefits, but instead can rely on social norms—specifically, the norms of the board, who, in Blair and Stout's view, are likely to be motivated by a reciprocity norm to protect the interests of minority shareholders who have placed trust in them. The prospect of legal liability, they claim, could crowd out such internal motivations and reduce trustworthy behavior.

Our BSM results provide reasons to question whether shareholders can rely on directors to protect them from opportunistic behavior of powerful CEOs or shareholders. Directors of firms with a powerful CEO or shareholder often can expect to benefit personally from supporting the transactions that the powerful actor wishes to conduct. Self-interest provides a strong motivation for directors to seek out BSM strategies to enable them to achieve their goal. And many options are likely to be available to them, including hiring agents (such as investment bankers) to advise them on the fairness of the transaction, and engaging in group-decisionmaking to dissipate responsibility (see Hamman et al., 2010; Bartling & Fishbacher, 2012; see Arlen & Tontrup, 2015a). Shareholder voting should further attenuate blame. They also can use their control over the decision-making environment to establish a norm under which people who provide outsized benefits to the company deserve outsized rewards.[25]

Our study provides reasons also to question that norms can do as much as some have hoped in curbing corporate crime, where incentives for misconduct are particularly high and both individuals and companies may have incentives to mute the constraining effect of legal norms (Arlen & Kornhauser, 2023; Feldman, 2014).[26] In circumstances in which there is potential personal or corporate benefit from noncompliance and BSM strategies are available, legal norm compliance is unlikely to be as robust in the real world, as we have suggested above.

3. Implications for Experimental Research

Our study, and our other BSM experiments (Tontrup & Sprigman, 2022, 2025; Arlen & Tontrup, 2015a; 2015b), also have important implications for experimental studies.

First and foremost, our study highlights a promising research agenda for experimental studies which is currently largely unexplored. Many behavioral deviations from rational choice operate to reduce self-interest by imposing an internal cost, such as through regret aversion, guilt or shame, on

[25] They can do so through the company's compensation and promotion system, and the internal messages to employees about what employee qualities are rewarded.

[26] There are multiple other reasons to contest this claim. One is that enforcement and sanctions are part of how society expresses its commitment to a legal injunction (Kahan, 1997; Arlen & Kornhauser, 2023). In addition, important categories of legal violations seek to enjoin harmful activity by people working for, and acting on behalf of, corporations. In such situations, the corporation, and not the law, is the institution that controls the direct levers that determine whether employees exist in a decision-making environment in which the law's injunctive norms are salient and established as the dominant social norm. Absent enforcement to ensure companies profit from compliance, companies can—knowingly or inadvertently—take multiple actions to undermine deterrence through expressive law in pursuit of profit (Arlen & Kornhauser, 2023).

the narrow pursuit of self-interest. We expect that these situations are ideally suited for individuals to use BSM.

Second, our results call for caution both in designing experimental studies and in making policy claims from those results in situations where the study is intended to test a behavioral mechanism that people could be motivated to mute or avoid through BSM. In such situations, it may be worth considering whether to add an explicit robustness test to assess whether any detected experimental results would stand up when subjects are afforded the opportunity to employ the BSM strategies most likely to be present in the real world. Policy conclusions also arguably should take into account the possibility that people's use of BSM strategies may render results of experiments that did not incorporate BSM opportunities less adjusted to real world situations where BSM opportunities are present.

4. How Far Does BSM Go?

This is not to say that people appear to always use BSM to remove the demands that social, legal or ethical norms place on them. Our claim is simply that there are real-world opportunities to use BSM techniques and that in our sample about a third of the mostly prosocial subjects chose to engage in BSM. However, prosocials engaging in BSM, combined with non-compliant proself types, constitute a majority of subjects who did not conform to the equal division norm.

We also do not claim that BSM strategies completely remove the disciplining constraints on self-interest arising from people's preference for norm compliance. Indeed, we observe that in our study, people who work to shift the norm reduced the demands of fairness norms, but did not eliminate them. These subjects continued, for the most part, to make positive transfers that evinced their preference to comply with a fairness norm, albeit the less-demanding effort-based norm. In Tontrup and Sprigman (2025) we find that subjects even though they delegate their allocation decision to an AI and significantly reduce their transfer violating the equal division norm, prosocials still make positive transfers

Yet, we do conclude that BSM can play a powerful role in dampening the effects of other-regarding preferences, even beyond the effects we report. BSM strategies can also operate in concert with other mechanisms that mute the demands of social norms. In real decision making, we expect behavior to be motivated by a mixture of BSM, wiggle room exploitation, and biased automatic processes like motivated reasoning to avoid normative constraints (discussed in Feldman, 2018). We see evidence of this joint impact in our study: While Allocators correctly anticipate that their work shifted the fairness norm in the eyes of Observers and act according to that norm, we also see them engage in wiggle room behavior: Allocators deliberately place their transfers in the norm range where they are less certain, and Observers are less certain whether they should consider the transfer fair. Finally, we see a suggestion of a self-serving bias (albeit small compared to the BSM norm-shifting effect) as the range of the fairness norm that Allocators indicate is shifted down in the favor relative to the Observers' true norm expectations.

We can imagine this combination of BSM and other mechanisms in many real-world scenarios: For example, principals may delegate a morally questionable task to an agent and correctly predict that this division of responsibility will reduce the accountability that impartial third parties will attribute to them for an unfair outcome. In addition, they may seek to obscure the true degree of

control they have over the agent to use this wiggle room to maintain a more positive social-image (Offer et al., 2024). Also, their perception of how much responsibility an actor actually bears and that others will attribute to them may be self-serving.

VI. Conclusion

This Article has shown the need to reassess the reliability of social and legal norms in guiding behavior, when individuals can reshape the moral context of their actions. Legal interventions and organizational rules that assume that norms and their effect on behavior are static, moral constraints rather than strategic variables, risk overestimating compliance. Moral and legal appeals are likely insufficient policy responses when actors are prepared to strategically shift rules in their favor at low personal and reputational costs. The norm-shifting mechanism we explored here is just one example of BSM. Other strategies that we intend to analyze or have already studied—such as the deliberate use of human and artificial agents to diffuse moral responsibility, or the installation of group decisions for the same goal—may similarly allow individuals to pursue self-interest while avoiding or at least reducing their moral costs. A detailed exploration of the reach and potential power of BSM strategies should be a priority for behavioral economics and legal scholarship.

References

Aldridge, S. J., et al. (2024). Uptake of COVID-19 vaccinations amongst 3,433,483 children and young people: Meta-analysis of UK prospective cohorts. *Nature Communications, 15,* Article 2363.

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

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal, 100(401),* 464–477.

Andreoni, J., & Bernheim, B. D. (2009). Social-image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica, 77(5),* 1607–1636.

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

Ariely, D., Bracha, A., & Meier, S. (2009). Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review, 99(1),* 544–555.

Arlen, J. (1998). The future of behavioral economics and the law. *Vanderbilt Law Review, 51(6),* 1765–1803.

Arlen, J. (2025). The compliance function. In J. Gordon & W.-G. Ringe (Eds.), *The Oxford Handbook of Corporate Law and Governance* (2nd ed.).

Arlen, J., & Kornhauser, L. (2023). Battle for our souls: A psychological justification for individual and corporate liability for organizational misconduct. *University of Illinois Law Review, 2023(2),* 673–721.

Arlen, J. & Kraakman, R. (1997). Controlling Corporate Misconduct: An Analysis of Corporate Liability Regimes**.** New York University. Law Review 72, 1997. 698, 1997.

Arlen, J., & Tontrup, S. (2015a). Does the endowment effect justify legal intervention? The debiasing effect of institutions. *Journal of Legal Studies, 44(1),* 143–168.

Arlen, J., & Tontrup, S. (2015b). Strategic bias shifting: Herding as a behaviorally rational response to regret aversion. *Journal of Legal Analysis, 7(2),* 517–544.

Babcock, L., Loewenstein, G., Issacharoff, S., & Camerer, C. F. (1995). Biased judgments of fairness in bargaining. *American Economic Review, 85*(5), 1337–1343.

Balliet, D., Parks, C. D., & Joireman, J. (2009). Social value orientation and cooperation in social dilemmas: A meta-analysis. *Group Processes & Intergroup Relations*, 12(4), 533–547.

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review, 3(3),* 193–209.

Bartling, B., & Fischbacher, U. (2012). Shifting the blame: On delegation and responsibility. *Review of Economic Studies, 79(1),* 67–87.

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

Battigalli, P., & Dufwenberg, M. (2022). Belief-dependent motivations and psychological game theory. *Journal of Economic Literature, 60(3),* 833–882.

Bazerman, M. H., & Tenbrunsel, A. E. (2011). *Blind spots: Why we fail to do what's right and what to do about it.* Princeton University Press.

Bénabou, R., & Tirole, J. (2011). Identity, morals, and taboos: Beliefs as assets. *Quarterly Journal of Economics, 126(2),* 805–855.

Bhattacharya, P. & Mollerstrom, J., Lucky to Work, *Review of Economics & Statistics* (2025)

Bilz, K., & Nadler, J. (2014). Law, moral attitudes, and behavioral change. In E. Zamir & D. Teichman (Eds.), *The Oxford Handbook of Behavioral Economics and the Law* (pp. 241–259). Oxford University Press.

Bicchieri, C. (2006). The grammar of society: The nature and dynamics of social norms. *Cambridge University Press.*

Bicchieri, C. (2017). Norms in the wild: How to diagnose, measure, and change social norms. *Oxford University Press.*

Blader, S. L., & Tyler, T. R. (2003). A four-component model of procedural justice: Defining the meaning of a "fair" process. *Personality and Social Psychology Bulletin, 29(6),* 747–758.

Blair, M. M., & Stout, L. (2001). Trust, Trustworthiness, and the Behavioral Foundations of Corporate Law. *U. Pa. L. Rev.* 149, 1735-1810.

Blake, P. R., McAuliffe, K., & Warneken, F. (2014). The developmental origins of fairness: Understanding fairness in children across cultures. *Journal of Experimental Psychology: General, 143*(2), 771–781.

Bochet, O., Page, T., & Putterman, L. (2006). Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior & Organization, 60(1),* 11–26.

Bogaert, S., Boone, C., & Declerck, C. (2008). Social value orientation and cooperation in social dilemmas: A review and conceptual model. *British Journal of Social Psychology, 47(3),* 453–480.

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

Bowles, S., & Gintis, H. (2011). *A cooperative species: Human reciprocity and its evolution*. Princeton University Press.

Bubb, R., & Pildes, R. H. (2014). How behavioral economics trims its sails and why. *Harvard Law Review, 127(5),* 1593–1678.

Camerer, C., & Thaler, R. (1995). Anomalies: Ultimatums, dictators, and manners. *Journal of Economic Perspectives, 9(2),* 209–219.

Capraro, V. (2018). Social versus moral preferences in the ultimatum game: A theoretical model and an experiment. *Theory and Decision, 85*(2), 193–214.

Capraro, V., & Rand, D. G. (2018). Do the right thing: Experimental evidence that preferences for moral behavior, rather than equity or efficiency per se, drive human prosociality. *Judgment and Decision Making, 13*(1), 99–111.

Cain, D. M., Loewenstein, G., & Moore, D. A. (2005). The dirt on coming clean: Perverse effects of disclosing conflicts of interest. *Journal of Legal Studies, 34(1),* 1–25.

Cappelen, C. and De Haan, T. (2023), How Much to Compensate the Unemployed: An Experimental Approach to Fair Unemployment Compensation" (Working Paper).

Celniker, J.B., Gregory, A., Koo, H. J., Piff, P. K., Ditto, P. H., and Shariff, A. F. (2023). The Moralization of Effort. *Journal of Experimental Psychology: General, 152(1),* 60-79.

Charness, G., Gneezy, U., & Henderson, A. (2018). Experimental methods: Measuring effort in economics experiments. *Journal of Economic Behavior & Organization, 149,* 74–87.

Charness, G., & Gneezy, U. (2008). What's in a name? Attention and effort in the dictator game. *Journal of Economic Behavior & Organization, 68(1–2),* 29–35.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics, 117(3),* 817–869.

Chen, Y., & Li, S. X. (2009). Group identity and social preferences. *American Economic Review, 99(1),* 431–457.

Cherry, T. L., Frykblom, P. & Shogren, J.F. (2002). Hardnose the Dictator, *American Economic Review, 92,* 1218-1221.

Coffman, C. (2011). Intermediation reduces punishment (and reward). *American Economic Journal: Microeconomics, 3(4),* 77–106.

Cooper, J. (2012). Cognitive dissonance theory. In P. A. M. van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology: Volume 1* (pp. 377–397). SAGE Publications.

Cooter, R. (1998). Expressive law and economics. *Journal of Legal Studies, 27,* 585–608.

Cooter, R. (2000). Do good laws make good citizens: An economic analysis of internalized norms. *Virginia Law Review, 86,* 1577–1601.

Cooter, R., & Eisenberg, M. A. (2001). Fairness, character, and efficiency in firms. *University of Pennsylvania Law Review, 149,* 1717–1732.

Cushman, F. (2008). Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108(2)*, 353–380.

Dana, J., Cain, D. M., & Dawes, R. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes, 100(2),* 193–201.

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

Dau-Schmidt, K. G. (1990). An economic analysis of the criminal law as a preference-shaping policy. *Duke Law Journal, 39(1),* 1–38.

De Cremer, D., & van Lange, P. A. M. (2001). Why prosocials exhibit greater cooperation than proselfs: The roles of social responsibility and reciprocity. *European Journal of Personality*, 15(S1), 5–18.

Depoorter, B., & Tontrup, S. (2024). Aspirational laws in action: A field experiment. *Law & Social Inquiry, 49(3)*, 1747–1782.

Dorff, M. B. (2003). Softening Pharaoh's heart: Harnessing altruistic theory and behavioral law and economics to rein in executive salaries. *Buffalo Law Review, 51,* 811–870.

Eckel, C. C., & Grossman, P. J. (1996). Altruism in Anonymous Dictator Games. *Games and Economic Behavior, 16(2)*, 181–191.

Engel, C. (2011). Dictator games: A meta-study. *Experimental Economics, 14(4),* 583–610.

Exadaktylos, F., Espín, A. M., & Brañas-Garza, P. (2013). Experimental subjects are not different. *Scientific Reports, 3,* Article 1213.

Faillo, M., Rizzolli, M., & Tontrup, S. (2019). Thou shalt not steal: Taking aversion with legal property claims. *Journal of Economic Psychology, 71,* 88–101.

Fahrenwaldt, A., Pesch, F., Fiedler, S., & Baumert, A. (2024). What's moral wiggle room? A theory specification. *Judgment and Decision Making, 19,* Article e17.

Falk, A. (2021). Facing yourself – A note on self-image. *Journal of Economic Behavior & Organization, 83(2),* 353–358.

Falk, A., Fehr, E., & Fischbacher, U. (2003). On the nature of fair behavior. *Economic Inquiry, 41(1),* 20–26.

Fehr, E., & Fischbacher, U. (2002). Why social preferences matter: The impact of non-selfish motives on competition, cooperation, and incentives. *Economic Journal, 112(478),* C1–C33.

Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences, 8*(4), 185–190.

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

Feldman, Y. (2014). Behavioral ethics meets behavioral law and economics. In E. Zamir & D. Teichman (Eds.), *The Oxford Handbook of Behavioral Economics and the Law* (Ch. 9). Oxford University Press.

Feldman, Y. (2018). *The law of good people: Challenging states' ability to regulate human behavior.* Cambridge University Press.

Feldman, Y., & Kaplan, Y. (2021). Preferences change and behavioral ethics: Can states create ethical people? *Theoretical Inquiries in Law, 22(1),* 85–101.

Fetchenhauer, D., & Dunning, D. (2012). Trust as a social and emotional act: Noneconomic considerations in trust behavior. *Journal of Economic Psychology, 33(3),* 686–694.

Fiedler, S., Glöckner, A., Nicklisch, A., & Dickert, S. (2013). Social value orientation and information search in social dilemmas: An eye-tracking analysis. *Organizational Behavior and Human Decision Processes, 120(2),* 272–284.

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters, 71(3),* 397–404.

Freddi, E. (2021). Do people avoid morally relevant information? Evidence from the refugee crisis. *The Review of Economics and Statistics, 103(4),* 605–620.

Frey, B. S., & Bohnet, I. (1999). Social distance and prosocial behavior: The case of generosity. *The Journal of Economic Behavior & Organization, 39(2),* 218–229.

Gerstenberg, T., & Lagnado, D. A. (2012). *When contributions make a difference: Explaining order effects in responsibility attributions. Psychonomic Bulletin & Review, 19(4)*, 729–736.

Götte, L., Stutzer, A., & Frey, B. M. (2010). Prosocial motivation and blood donations: A survey of the empirical literature. *Transfusion Medicine and Hemotherapy, 37(3),* 149–154.

Greenfield, K. (2001). Using Behavioral Economics to Show the Power and Efficiency of Corporate Law as a Regulatory Tool, *U.C. Davis L. Review,* 35, 581-644.

Gross, J., & Vostroknutov, A. (2022). Why do people follow social norms? *Current Opinion in Psychology, 44*, 1–6.

Grossman, Z. (2014). Strategic ignorance and the robustness of social preferences. *Management Science, 60(11),* 2659–2665.

Grossman, Z. (2015). Self-signaling and social-signaling in giving. *Journal of Economic Behavior & Organization, 117*, 26–39.

Grossman, Z., & Oexl, R. (2012). Delegating to a powerless intermediary: Does it reduce punishment? *Experimental Economics, 16(2),* 306–322.

Grossman, Z., & van der Weele, J. J. (2017). Self-image and willful ignorance in social decisions. *Journal of the European Economic Association, 15(1),* 173–217.

Hamman, J. R., Loewenstein, G., & Weber, R. A. (2010). Self-interest through delegation: An additional rationale for the principal-agent relationship. *American Economic Review, 100(4),* 1826–1846.

Harris, L. T., & Fiske, S. T. (2011). Dehumanized perception: A psychological means to facilitate atrocities, torture, and genocide? *Journal of Psychology, 219(3)*, 175-181.

Haesevoets, T., Reinders Folmer, C., & van Hiel, A. (2014). More money, more trust? Target and observer differences in the effectiveness of financial overcompensation to restore trust. *Psychologica Belgica*, 54(4), 389-394.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of Homo Oeconomicus: Behavioral experiments in 15 small-scale societies. *American Economic Review, Papers and Proceedings*, 91(2), 73-78.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N. S., Hill, K., Gil-White, F., Gurven, M., Marlowe, F. W., Patton, J. Q., & Tracer, D. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, 28(6), 795-815**.**

Hill, A. (2015). Does delegation undermine accountability? Experimental evidence on the relationship between blame shifting and control. *Journal of Empirical Legal Studies, 12(2),* 311–334.

Hoffman, E., & Spitzer, M. (1985). Entitlements, rights, and fairness: An experimental examination of subjects' concepts of distributive justice. *Journal of Legal Studies, 14(2),* 259–297.

Hoffman, E., McCabe, K., Shachat, K. & Smith, V. (1994). Preferences, Property Rights, and Anonymity in Bargaining Games. *Games and Economic Behavior*, 7(3), 346-80.

Hoffman, E., McCabe, K., & Smith, V. L. (1996). Social distance and other-regarding behavior in dictator games. *American Economic Review, 86(3),* 653–660.

Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science, 345(6202),* 1340–1343.

Hollander-Blumoff, R. (2017). Social value orientation and the law. *William & Mary Law Review, 59(2),* 475–539.

Johnson, N. D., & Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology, 32(5),* 865–889.

Jolls, C., Sunstein, C. R., & Thaler, R. H. (1998). A behavioral approach to law and economics. *Stanford Law Review, 50(5),* 1471–1550.

Kahan, D. M. (1997). Social influence, social meaning, and deterrence. *Virginia Law Review, 83,* 349–395.

Kanagaretnam, K., Mestelman, S., Nainar, K., & Shehata, M. (2009). The impact of social value orientation and risk attitudes on trust and reciprocity. *Journal of Economic Psychology*, 30(3), 368-380.

Kanngiesser, P., & Warneken, F. (2012). Young children consider merit when sharing resources with others. *PLOS ONE, 7(8),* e43979.

Karagözoğlu, E. & Urhan, U.B. (2017). The effect of stake size in experimental bargaining and distribution games: A survey *Group Decision and Negotiation*, 26 (2) , 285-325.

Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association, 11(3),* 495–524.

Lamont, J. (1994). The concept of desert in distributive justice. *Philosophical Quarterly, 44(174),* 45–64.

Langevoort, D. C. (2018). Behavioral ethics, behavioral compliance. In J. Arlen (Ed.), *Research Handbook on Corporate Crime and Financial Misdealing* (pp. 261-285). Edward Elgar Publishing.

Larney, A., Rotella, A., & Barclay, P. (2019). Stake size effects in ultimatum game and dictator game offers: A meta-analysis. *Organizational Behavior and Human Decision Processes, 151*, 61–72.

Lazear, E. P., Malmendier, U., & Weber, R. A. (2012). Sorting in experiments with application to social preferences. *American Economic Journal: Applied Economics, 4(1),* 136–163.

Lin, P.-H., Brown, A. L., Imai, T., & Wang, J. T.-Y. (2020). Evidence of general economic principles of bargaining and trade from 2,000 classroom experiments. *Nature Human Behaviour, 4(9),* 1–11.

List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy, 115(3),* 482–493.

Lu, X., Kornhauser, L., & Tontrup, S. (2025). Managing crowding out effects: When prosocial behaviors can be incentivized. *NYU Working Paper.*

Matthey, A., & Regner, T. (2014). More than outcomes: The role of self-image in other-regarding behavior. *Jena Economic Research Papers* (Working Paper 2014–036).

Mazar, N., & Zhong, C. B. (2010). Do green products make us better people? *Psychological Science, 21(4),* 494–498.

McAdams, R. (1997). The origin, development, and regulation of norms. *Michigan Law Review, 96,* 338–433.

Mischkowski, D. Stone, R. & Stremitzer, A. (2019). *Promises, Expectations, and Social Cooperation. Journal of Law & Economics*, 62 (4), 687–712.

Momsen, K., & Ohndorf, M. (2020). When do people exploit moral wiggle room? An experimental analysis of information avoidance in a market setup. *Ecological Economics, 169,* 106479.

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. J. J. (2011). Measuring social value orientation. *Judgment and Decision Making, 6(8),* 771–781.

Offer, K., Rahwan, Z., & Hertwig, R. (2024). Foucault's error: The power of not knowing. *European Review of Social Psychology,* 1–36.

Offerman, T., Sonnemans, J., & Schram, A. (1996). Value orientations, expectations, and voluntary contributions in public goods. *The Economic Journal, 106(437),* 817–845.

Oliver, A. (2024). If you've earned it, you deserve it: Ultimatums, with Lego. *Behavioural Public Policy, 8,* 395–402.

Ogawa, K., Takemoto, T., Takahashi, H., & Suzuki, A. (2010). The belief that others think effort should be rewarded: Experimental evidence in dictator games, *Economics and Management Working Paper No. 2010-E04 Kyoto Sangyo University*.

Oxoby, R. J. and J. Spraggon (2008), Mine and yours: Property rights in dictator games, *Journal of Economic Behavior & Organization*, 65: 703–13.

Ostrom, E. (2000). Collective action and the evolution of social norms. *Journal of Economic Perspectives, 14*(3), 137–158.

Petersen, M. B., Sznycer, D., Cosmides, L., & Tooby, J. (2012). Who Deserves Help? Evolutionary Psychology, Social Emotions, and Public Opinion about Welfare. *Political Psychology, 33(3)*, 395–418.

Pletzer, J. L., Balliet, D., Joireman, J., Kuhlman, D. M., Voelpel, S. C., & van Lange, P. A. M. (2018). Social value orientation, expectations, and cooperation in social dilemmas: A meta-analysis. *European Journal of Personality*, 32(1), 62-83.

Regner, T. (2021). What's behind image? Toward a better understanding of image-driven behavior. *Frontiers in Psychology, 12,* 614575.

Schier, U. K., Ockenfels, A., & Hofmann, W. (2016). Moral values and increasing stakes in a dictator game. *Journal of Economic Psychology, 56,* 107–115.

Schmidtz, D. (2006). *Elements of justice.* Cambridge University Press.

Shaw, A., & Olson, K. R. (2012). Children discard a resource to avoid inequity. *Journal of Experimental Psychology: General, 141(2)*, 382–395.

Sher, G. (1987). *Desert.* Princeton University Press.

Spiekermann, K., & Weiss, A. (2016). Objective and subjective compliance: A norm-based explanation of "moral wiggle room. *Games and Economic Behavior, 96*, 170–183.

Starmans, C., Sheskin, M., & Bloom, P. (2017). Why people prefer unequal societies. *Nature Human Behaviour, 1(4),* 0082.

Stout, L. A. (2002). In praise of procedure: An economic and behavioral defense of *Smith v. van Gorkom* and the business judgment rule. *Northwestern University Law Review, 96,* 675–684.

Stout, L. A. (2011). *Cultivating conscience: How good laws make good people.* Princeton University Press.

Sunstein, C. R. (1996a). On the expressive function of law. *University of Pennsylvania Law Review, 144(5),* 2021–2053.

Sunstein, C. R. (1996b). Social norms and social roles. *Columbia Law Review, 96(4),* 903–968.

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology, 58,* 345–372.

Tontrup, S., & Sprigman, C. J. (2022). Self-nudging contracts and the positive effects of autonomy — Analyzing the prospect of behavioral self-management. *Journal of Empirical Legal Studies, 19*(3), 594–676.

Tontrup, S. & Sprigman, C. J. (2025). *Strategic Delegation of Moral Decisions to AI*. SSRN Working Paper.

Tyler, T. R. (2021). *Why people obey the law* (3rd ed.). Princeton University Press.

Umer, H., Kurosaki, T., & Iwasaki, I. (2022). Unearned endowment and charity recipient lead to higher donations: A meta-analysis of the dictator game lab experiments. *Journal of Behavioral and Experimental Economics, 97,* 101825.

van den Bos, K., van Dijk, E., Westenberg, M., & Yperen, N. W. V. (2009). On the psychology of cooperation in social dilemmas: The influence of SVO. *Journal of Experimental Social Psychology, 45(6),* 1230–1235.

van Dijk, F., Sonnemans, J., & van Winden, F. (2002). Social ties in a public good experiment. *Journal of Public Economics*, 85(2), 275-299.

van Lange, P. A. M. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology, 77(2),* 337–349.

van Lange, P. A. M., Bekkers, R., Schuyt, T. N. M., & van Vugt, M. (2007). From games to giving: Social value orientation predicts donations to noble causes. *Basic and Applied Social Psychology*, 29(4), 375-384.

Vanberg, C. (2008). Why do people keep their promises? An experimental test of two explanations. *Econometrica, 76(6),* 1467–1480.

Yamagishi, T., Mifune, N., Li, Y., Shinada, M., Hashimoto, H., Horita, Y., Miura, A., Inukai, K., Tanida, S., Kiyonari, T., Takagishi, H., & Simunovic, D. (2013). Is behavioral pro-sociality game-specific? Pro-social preference and expectations of pro-sociality. *Journal of Economic Psychology*, 34, 13-26.

Zhong, C. B., & Liljenquist, K. (2006). Washing away your sins: Threatened morality and physical cleansing. *Science, 313(5792),* 1451–1452.

Appendix

I. Methods

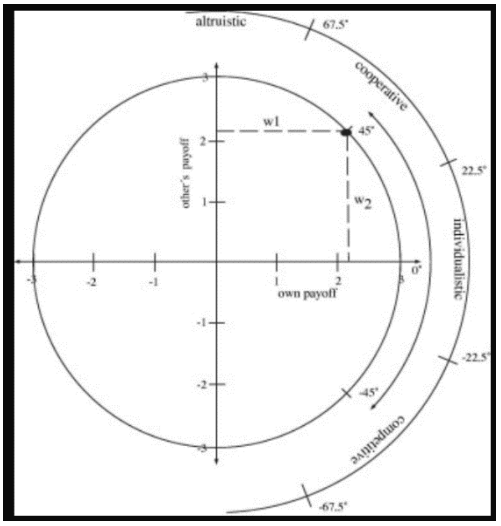
A. Social Value Orientation

We elicit subjects' social value orientation (SVO), a measure of how much an individual cares about other players' outcome in relation to her own.[37] The ring measure has been shown to be predictive of cooperative behavior in many studies and across different games and scenarios: for example, in public goods games studied in De Cremer & van Lange (2001) and Fiedler et al. (2013), in one-shot and repeated prisoner's dilemma tasks in Balliet et al. (2009), Offerman et al. (1996) and Pletzer et al. (2018), in investment games in Kanagaretnam et al. (2009), trust games are studied in Yamagishi et al. (2013) and Haesevoets et al. (2014), and in field studies by van Lange et al. (2007).

We employ the ring measure developed by van Dijk, Sonnemans and van Winden (2002) that follows the basic design of Liebrand and McClintok, but is incentivized and adds further scenarios. The measure asks subjects to choose between 32 pairs of allocations for themselves and a randomly assigned partner. For instance, the first pair asks subjects to choose between allocation A = (0, 500) and allocation B = (305, 397). Choosing A would result in a payment of x=0 points (convertible into money at the rate of 1 point to €0.001 for all choices) for the subject and y=500 points for her partner; choosing B would allocate x=305 points to the subject and y=397 points to her partner. See Appendix B for all 32 pairs of these allocations. Each of the 32 allocations can be represented as a vector using Cartesian coordinates—payment-to-self on the x-axis and payment-to-partner on the y-axis. To obtain the social value orientation of the subject we sum up these 32 vectors and calculate

the angle that the resultant vector makes with the horizontal axis. The length of the resultant vector divided by 1000 provides us with a score (between 0 and 1) that informs us about how consistently the subject's choices fit a particular social type. The SVO classifies individuals as individualistic (focused on their own payoff), cooperative (concerned with the sum of their own and their partner's joint payoff), altruistic (focused on their partner's payoff), and competitive (aiming to increase the difference between their own and their partner's payoff). This allows us to classify our subjects into competitive (between -67.5 and -22.5 degrees), individualistic (between -25 and +22.5 degrees), cooperative (between 22.5 and 67.5 degrees) and altruistic types (between 67.5 and 112.5 degrees), as depicted in the graph below.[38]

tFigure A1. Illustration of the SVO Ring Measure



The construction of the ring measure is important for understanding our hypothesis. The SVO is a continuum; most subjects are not purely selfish or even competitive, nor are they purely prosocial. They may choose in one allocation decision to benefit their own interests if the benefit is large and outweighs their other-regarding preferences. On the other hand, if their personal gain is smaller compared to the gain of the other, they may forgo their own gain and make a prosocial choice. Thus, when we distinguish between proself and prosocial types, it should be understood that most proself individuals will also have other-regarding preferences of some degree, while prosocial individuals will also act in their self-interest in some situations, while nonetheless having other-regarding preferences that are relatively stronger overall. The presence in different degrees of both kinds of preference in many individuals explains why we expect both prosocial types and proself types to take on the work task in order to reduce their transfers and moral costs. However, we expect the effect to be stronger for prosocial (cooperative and altruistic) types than for proself types (individualistic and competitive), because their more pronounced prosociality suggests that they would face higher moral costs for deviating from a fairness norm to pursue their self-interest to the detriment of someone else, and thus they should have a stronger incentive to work. The SVO score should primarily correlate with subjects' self-image concerns, since we implement a strict anonymity protocol throughout the SVO and there is no Observer to judge participants' SVO decisions. However, since most individuals who are intrinsically concerned about their moral self-image will also be concerned about their

social-image, the SVO should also approximate social-image concerns (except for individuals who only want to appear prosocial to others).

B. The Work Task

R 1	8	1	7	1	1	6	1	3	5	1	4	4	7	1	0	5	1	5	1	6
R 2	0	3	2	0	9	8	3	1	5	8	7	8	7	8	3	4	1	4	2	1
R 3	6	7	9	6	1	4	6	8	6	5	5	3	1	1	3	1	9	2	6	1
R 4	8	6	5	5	7	0	9	6	6	1	3	2	6	5	1	1	1	1	1	3
R 5	1	9	1	9	8	7	1	5	4	1	1	5	1	7	1	1	0	1	1	2
R 6	1	1	1	9	6	0	2	1	9	0	8	6	0	0	8	7	5	3	2	1
R 7	8	7	3	0	5	2	1	3	4	1	7	5	1	4	1	8	6	1	2	3
R 8	1	6	2	5	2	8	4	2	1	3	5	1	6	1	8	1	7	6	1	1
R 9	1	5	3	4	6	1	4	0	5	1	0	4	1	1	9	1	3	4	8	4
R 10	9	1	5	3	7	1	1	9	5	6	6	1	1	1	7	8	5	8	0	1

C. Control Questions

Q1. If you transfer €x of your €12 endowment to the Recipient, how many Euros would you keep for yourself and how many would the Recipient receive?

You keep [€] Recipient receives [€]

Q2. Can you increase your endowment by completing the work task?

[yes] [no]

Q3. Is your partner in the Social Value Orientation Measure the same as in the main experiment? [yes] [no]

Q4. Can we link the choices you make in the experiment with your identity?

[yes] [no]

II. Results

A. Norm Alignment

To analyze norm alignment, we focus on the minimum transfers that Allocators expected Observers to consider fair and the minimum transfers that Observers actually identified as fair. We construct two variables, one for Allocators and one for Observers. The first variable conditions Allocators' lower-bound norm expectations on their actual work decision: for Allocators who worked, we use the minimum transfer they expected Observers will consider fair, given that they worked; for those who did not work, we use their expectation of the minimum fair transfer, given that they did not work. We then construct a corresponding variable for Observers—the minimum transfer they considered fair (at 50% confidence), conditional on whether their matched Allocator worked or did not work.

We then compute the difference between these two variables by subtracting the Observer's minimum fair transfer from the Allocator's expected minimum fair transfer. A negative value

indicates that the Allocator's norm expectation is lower than what the Observer actually considers fair, whereas a value of zero or greater indicates that the Allocator's expectation meets the Observer's fairness perception. On this basis, we code a binary variable that takes the value 0 if the Allocator's expectation falls short and 1 if it meets or exceeds the minimum of what the Observer considers fair. We then calculate the percentage of cases coded as 1, which we refer to as cases of norm alignment.

To assess whether norm alignment occurs more often than would be expected by chance, we perform a binomial test. We use $p= 0.5$ as a statistical benchmark representing a situation with no tendency in either direction—that is, aligned and non-aligned pairs are equally frequent—and test whether the observed alignment rate exceeds this value.

Hypothesis: ***The alignment of expectations in our Allocator-Observer pairs is significantly higher than p=0.5***

The data support the hypothesis: Across the two Treatments 75.4% of the Allocators held lower-bound norm expectations that did not fall short of those of the Observers—a proportion significantly higher than the benchmark of 0.5 ($p < 0.01$). The high alignment rate also appears to suggest that most Allocators do not seem to exhibit a self-serving bias.

To further investigate whether an effort-based fairness norm has emerged, we compare for every Allocator who completed the effort task, their stated lower-bound estimate of what they think Observers might consider a fair transfer with each Observer's norm expectation of the uncertain range and compute the percentage of Observers whose normative demands are met; averaging these norm alignments within a treatment estimates how widely Allocators and Observers accept the shift to an effort-based norm. We find that the 40 workers in Treatment 1 satisfy on average 65% of Observer norm expectations. The workers in Treatment 2 satisfy 66.8%. Pooling the 70 workers across both treatments yields an overall alignment of 66.2%. Thus, almost two Allocators in three meet the norm expectations of Observers, indicating that the effort-based norm enjoys relatively broad—though not complete—acceptance. To confirm this statistically, we treat each workers' individual alignment percentage as one independent observation and test whether the average proportion exceeds fifty percent. A Wilcoxon signed-rank test rejects the 50-percent benchmark in both treatments and in the pooled sample (all $p < 0.01$), supporting the conclusion that an effort-based norm has taken hold.

B. Allocators Exploit Wiggle Room

As we have seen above, Allocators understand that the new effort-based norm they created through their work is more ambiguous than the equal distribution norm which would have applied otherwise. While we have seen that Allocators comply with their expectations about Observers' views of what fair transfer requires, we predicted they would make use of the "moral wiggle room" created by the shift to a broader and more uncertain norm. As explained, whereas absent work Allocators and Observers converge on a discrete social norm of equal distribution, it is more difficult to have the same clear shared understanding regarding what is the fair transfer when Allocators

undertook the work task. The transfer amounts that Allocators are certain or somewhat confident that Observers will view as fair range from €2.5 to €5.0 across both treatments. Similarly, Observers' norm perceptions of a fair transfer range from €3.1 to €5.3. We assume that Allocators aim to exploit the wiggle room they created by shifting to a more uncertain and broader effort norm making transfers at the lower end of the range that they expect Observers to believe would be fair.

We posit

Hypothesis D_5: ***A significant number of Allocators will utilize moral wiggle room.***

Consistent with our hypothesis, Allocators who worked selected a transfer within the range of their norm expectations, but at the level they were only partially sure the Observer would consider fair. Of those who worked, 72.6% chose a transfer amount that fell within the range where they were only somewhat confident the Observers would consider it fair. In contrast, the group, 27.4%, that selected a transfer they were certain the Observers would consider a fair transfer, was significantly smaller, providing strong support for $\mathbf{D_5}$ ($\chi^2 = 8.63$, $p < 0.01$).

This suggests that Allocators leverage the ambiguity of the new norm and make a strategic decision that navigates between earning a higher payoff and the risk of suffering social-image costs in turn.

C. Self-Image and Social-Image Motivation

We have shown in the article that both social-image and self-image concerns drive BSM in our study. We now isolate these two motivations to demonstrate that each independently contributes to participants' behavior. We do so in two ways: first, by examining their separate influence on the decision to work; and second, by analyzing their influence on norm compliance—i.e., the likelihood that Allocators meet their own expectations of what Observers would consider a fair transfer.

1. Influence on the Decision to Work

To disentangle the effects of self-image and social-image concerns on work choices, we estimate two complementary regressions. The first restricts the sample to the Self-Image Game, where the coefficient on the treatment dummy captures the effect of self-image concerns relative to the control group. The second is a panel logistic regression that includes a game dummy distinguishing between the Self-Image and Social-Image Games, a treatment dummy capturing the combined effect of both treatments relative to the control condition, and an interaction term between treatment and game. In this specification, the main effect of treatment reflects the combined motivational influence of both treatments, while the interaction term subtracts the component specific to the Social-Image Game—thereby isolating the incremental effect of social-image concerns.

The regression restricted to the Self-Image Game shows a significant treatment effect (odds ratio = 10.87, $p < 0.001$). In the full panel regression, the interaction term is –1.40 ($p = 0.031$), indicating that the Social-Image Game has a statistically significant incremental effect over and above the influence of self-image alone. These results suggest that both self-image and social-image concerns independently motivate Allocators to engage in BSM by choosing to complete the effort task.

2. Influence on Norm Compliance.

We now assess compliance with what participants believe Observers would consider fair, using two complementary tests. First, we conduct one-sample z-tests of proportions within each game to assess whether the share of compliant Allocators exceeds a 50% benchmark—that is, whether participants comply with their own norm expectations more often than would occur by random choice. We see compliance significantly exceeding chance in both games: In the Self-Image Game, compliance is 0.65 (z= 4.24, p< 0.001), and in the Social-Image Game it is 0.77 (z= 7.64, p< 0.001). Thus, even in full privacy, Allocators comply significantly more often than random choice would predict, and observability further increases this tendency.

Second, we compare compliance across the two games within subjects. Because the same 200 participants make both decisions, we use McNemar's paired-proportions test to evaluate whether the presence of an Observer increases compliance. The test yields z= –2.64 (p= 0.02), an increase in compliance when decisions are observed.

Taken together, these results demonstrate that self-image concerns can, working alone, drive people to use BSM when they want to increase their share of the pie, but the addition of social-image concerns makes even more individuals engage in BSM and work.

III. Instructions

Instructions for the Article: Behavioral Self-Management and the Strategic Shifting of Fairness Norms

by Stephan Tontrup, Jennifer Arlen & Christopher Jon Sprigman

I. Introduction

Dear Participant,
Thank you for participating in our study and for completing the consent form! All necessary instructions for the experiment will be displayed on your screen. You will be paid via PayPal after completing the experiment.

Important: You can only use your login key once. If you abort the study before completing the experiment in full, you cannot continue and will receive no payment.
In the following instructions, all possible decisions you can make during the experiment and their consequences are described. Please note that your payment depends on the decisions you make [Note: except in the Control Treatment]. It is therefore very important that you read these instructions carefully.

A. Anonymity and Duration

All participants remain strictly anonymous. As already mentioned in the invitation to the experiment, the experiment takes approximately 30 minutes.

To ensure your anonymity, we have asked you to use an email address for participation that is unknown to us, contains no identifiers (e.g., no name), and is registered with PayPal for payment. We have provided you with instructions for creating alias email addresses.

Important: If you use an email address with identifiers, we cannot use your data and you will receive no payment. If you do not have an appropriate email address and still wish to participate, please log out now before beginning the study, create an email address according to our instructions, and then log back in.

Additionally, we have asked you to use a VPN for participation in the experiment. Our system does not collect IP addresses, so you are anonymous to the experimenters. However, the VPN ensures that you yourself can be certain that you are completely anonymous. Neither the experimenters (nor anyone else) can establish any connection between you and your decisions in the experiment. You remain completely anonymous. After completion of the study, the experimenters will contact you and inform you about the study and its objectives.

*****

B. Treatments [for Self-Image Game identical for Treatment 1 & Treatment 2; the order of Self-Image and Social-Image games was randomized]

Participants in this experiment are randomly assigned a role: you are either an Allocator or a Recipient. You have been randomly selected as an Allocator. A second participant has been randomly assigned the role of "Recipient". In your role as Allocator, you receive 12 euros. You decide how to divide the 12 euros between yourself and the Recipient. The Recipient has a passive role and receives no endowment in euros. You can choose any distribution between yourself and the Recipient (from 0-12 euros for yourself and 0-12 euros for the Recipient). You keep the amount you assign to yourself.
At the same time, you yourself are assigned to an Allocator as a Recipient. This Allocator is not the participant who is assigned to you as Recipient, but a third participant. As a Recipient, you additionally receive the amount that this Allocator allocates to you.

Work Task [Note: not in Baseline, where no work task is offered]

Before making your allocation decision, you can voluntarily complete a task. Completing the task is not difficult and can be managed by anyone, but it requires time and attention. Whether you complete the task is your decision. Completing the task has no effect on your endowment of 12 euros: neither you nor the Recipient receive an additional amount as a result. The Recipient cannot complete a task themselves.
Here is an example of the task you can complete. On the screen you will see digits from "1" to "9". You will be asked to count all "1" digits in the box.
If you enter an incorrect solution, you must repeat the task until you enter the correct solution. To prevent it from being effective to find the solution by trial and error, you must wait 30 seconds after each incorrect entry before you can make another entry. A timer runs showing you by when you must have made an entry so that the experiment is not terminated. If you abort the experiment, you will receive no payment.

Below you see an example of a number box. Your task is to count how often the number "1" appears in the box. To complete the task as a trial, you can make an entry at the bottom of the screen indicating how many ones you believe appear in this table. You will then receive feedback.

A	5	3	5	2	1	4	2	6	4	7	3	5	1	7	5	8	9	3	9	7
B	2	4	3	6	1	3	5	0	1	7	8	5	0	2	6	0	9	9	2	2
C	2	7	1	9	0	6	9	3	6	5	3	8	2	4	2	6	4	8	9	1
D	3	4	5	2	7	1	5	2	9	5	0	7	8	8	4	6	2	2	1	0
E	2	0	6	8	7	7	2	5	1	7	3	0	5	3	7	1	9	0	1	6
F	5	3	1	6	4	8	1	5	2	9	8	0	5	0	3	9	4	5	2	7
G	1	2	6	1	7	8	3	0	4	7	2	8	3	8	6	1	7	3	8	4
H	1	5	2	8	5	0	1	6	3	8	3	9	0	0	4	7	2	8	2	7
I	5	5	2	8	4	9	1	6	2	8	3	0	2	4	1	6	2	7	4	9
J	8	3	7	2	8	5	9	3	7	1	8	3	7	5	0	1	5	3	8	1

Number of "1" [ ]

C. Control Questions

1. If you transfer x euros from your 12 euros, how many euros do you keep for yourself and how many does the Recipient receive?
You [12-x] Recipient [x]

2. How many additional euros do you receive if you complete the work task correctly? [not in Baseline]
[0] euros

3. Does the Recipient know that your decision to work is voluntary?
[0] Yes [0] No [correct answer in Treatment 1=no, in Treatment 2=yes; not in Baseline]

4. Decision about the Work Task [not in Baseline]

You can now decide whether you would like to complete the work task or not.
Would you like to complete the work task?
[] Yes [] No

5. Decision about the Allocation

Now we ask you to decide how you would like to divide the 12 euros between yourself and the Recipient. Please indicate the division in euros below.
[] You [] Recipient

Second Game – [Surprise Restart – Social-Image Game; which game is played first is randomized]

[Treatment 2 – instructions for Treatment 1 follow below]

We now present to you again the allocation experiment that you just completed. You will therefore again allocate 12 euros between yourself and the Recipient and can decide whether you want to complete the work task.
[varies between Treatment 2 & Treatment 1] *However, there is a change: A third participant - a neutral Observer - also takes part in the experiment. This Observer sees your transfer decision. Furthermore, the Observer is comprehensively informed about the work task: The Observer sees the ones-counting task and learns that completing it is voluntary for you, that neither you nor the Recipient are paid for completing it (your endowment therefore remains the same), that the Recipient cannot work themselves, and whether you have completed the task or not.*
Additionally, the Observer has an endowment of 50 euro-cents, which they can pass on to you in full or in half; the amount they do not pass to you, is lost. After the experiment, you will be informed about the Observer's decision and paid accordingly.
Now you can first decide again whether you want to complete the work task or not. Here is the task once more:

4. Decision about Work Task

Please now make your decisions. Do you want to complete the task?

[] Yes [] No

5. Decision about the Allocation

Now we ask you to decide how you would like to divide the 12 euros between yourself and the Recipient. Please indicate the division in euros below.
[] You [] Recipient

[Treatment 1]

We now present to you again the experiment that you just completed. You will therefore again divide 12 euros between yourself and the Recipient and can decide whether you want to complete the work task.
[varies between Treatment 1 & Treatment 2] *However, there is a change: A third participant - an Observer - also takes part in the experiment. This Observer sees your transfer decision. The Observer receives the following information about the work task: If you have completed the ones-counting task, the Observer learns that you have completed the task, that you are not paid for it (your endowment therefore remains the same) and that the Recipient cannot work themselves - but not that completing it was voluntary for you. If you have not completed the task, the Observer learns nothing about its existence; from the Observer's perspective, the experiment consists only of the transfer decision without a work task.*
Additionally, the Observer has an endowment of 50 euro-cents, which they can pass on to you; if they do not pass the amount to you, it is lost. After the experiment, you will be informed about the third participant's decision and paid accordingly.

II. Control-Treatment (non-monetary allocation)

……

2. Experiment

Participants in this experiment are randomly assigned a role: you are either an Allocator or a Recipient. You have been randomly selected as an Allocator. A second participant has been randomly assigned the role of Recipient. In your role as Allocator, you receive 12 non-monetary points. You decide how to divide the 12 points between yourself and the Recipient. The Recipient has a passive role and receives no endowment of points. You can choose any distribution between yourself and the Recipient (from 0-12 points for yourself and 0-12 points for the Recipient). You keep the points that you assign to yourself.
At the same time, you yourself are assigned to an Allocator as a Recipient. This Allocator is not the participant who is assigned to you as Recipient, but rather a third participant. As a Recipient, you additionally receive the points that this Allocator allocates to you.

Work Task

Before making your allocation decision, you can voluntarily complete a task. Completing the task is not difficult and can be managed by anyone, but it requires time and attention. Whether you complete the task is your decision. Completing the task has no effect on your endowment of 12 non-monetary points: neither you nor the Recipient receive additional points as a result. The Recipient cannot complete a task.
Here is an example of the task you can complete. On the screen you will see digits from "1" to "9". You will be asked to count all "1" digits in the box.
If you enter an incorrect solution, you must repeat the task until you enter the correct solution. To prevent it from being effective to find the solution by trial and error, you must wait 30 seconds after each incorrect entry before you can make another entry. A timer runs showing you by when you must have made an entry so that the experiment is not terminated. If you abort the experiment, you will receive no payment.

Example Task

Below you see an example of a number box. Your task is to count how often the number "1" appears in the box. To complete the task as a trial, you can make an entry at the bottom of the screen indicating how many ones appear in this table. You will then receive feedback.

[same number box appears as above]

3. Control Questions

1. If you transfer x points from your 12 points, how many points do you keep for yourself and how many does the Recipient receive?
You [12-x] Recipient [x]
2. How many additional euros do you receive if you complete the work task correctly? [not in Baseline]

[0] euros
3. Does the Recipient know that your decision to work is voluntary?
[0] Yes [0] No [correct answer in Treatment 1=no, in Treatment 2=yes; not in Baseline]

4. Decision about the Work Task [not in Baseline]

You can now decide whether you would like to complete the work task or not.
Would you like to complete the work task?
[] Yes [] No

5. Decision about the Allocation

Now we ask you to decide how you would like to divide the 12 points between yourself and the Recipient. Please indicate the division of points below.
[] You [] Recipient

Surprise Restart – Social-Image Game [which game is played first is randomized; in the Control Treatment, the instructions of Treatment 2 are used]

We now present to you again the allocation experiment that you just completed. You will therefore again allocate 12 non-monetary points between yourself and the Recipient and can decide whether you want to complete the work task.
However, there is a change: A third participant - a neutral Observer - also takes part in the experiment. This Observer sees your transfer decision. Furthermore, the Observer is comprehensively informed about the work task: The Observer sees the ones-counting task and learns that completing it is voluntary for you, that you are not paid for it (your endowment therefore remains the same), that the Recipient cannot work themselves, and whether you have completed the task or not.
After the experiment, you will be informed about the Observer's decision and paid accordingly.

Control Question.

Does the Observer know that your decision to work is voluntary?
[0] Yes [0] No [correct answer in Treatment 1=no, in Treatment 2=yes; not in Baseline]
Now you can first decide again whether you want to complete the work task or not. Here is the task once more...

III. Measurement Instruments and Further Data Collection

A. Allocators' Expectations of Whether Observer Considers Transfer Fair [Social-Image Game]

For each possible transfer amount from 0 € to 12 € (in 1-€ steps), please indicate how confident you are that the Observer considers this transfer amount fair. "Not confident" suggests that you do not think that the Observer will view the transfer as fair. Select exactly one of the following levels per amount:

100% = completely confident
50% = rather confident
0% = not confident

Payment:

You receive a €1 endowment. The computer randomly picks a stake from this endowment. The stake is multiplied based on whether you are right with your assessment, i.e. that the Observer indeed considers the transfer fair. You keep the amount from the €1 endowment that was not picked as the stake. There is a payment rule that ensures it is always optimal for you to state your true assessment.

100%: Right → 3×stake / Wrong → €0

50%: Right → 2.5×stake / Wrong → 0.5×stake

0%: Right → €0 / Wrong → 3×stake

Your total payment for your assessment is = rest of endowment + payout on stake.

Here is a simple Example. Assume the randomly drawn stake is = €0.40.

> You expect "Fair" at 100%, Observer indicates "Fair" → €0.60 (the rest of the €1 endowment) + €1.20 (3x the drawn stake)= €1.80; Observer says "Unfair"= €0.60 (rest of endowment; the stake is lost)
>
> You expect "Fair" at 50%, Observer indicates "Fair" → €0.60 + €1.00 = €1.60; Observer says "Not Fair" → €0.60 + €0.20 = €0.80
>
> You expect "Not Fair" at 0%, Observer indicates "Not fair" → €0.60 + €1.20 = €1.80; Observer indicates "Fair" → €0.60

Now please make your assessment for each potential transfer amount:

1. You transfer 0 euros to the Recipient. How confident are you that the Observer would consider this transfer fair?

I am ... [] in this assessment

2. You transfer 1 euro to the Recipient. How confident are you that the Observer would consider this transfer fair?

I am ... [] in this assessment

3. You transfer 2 euros to the Recipient. How confident are you that the Observer would consider this transfer fair?

I am ... [] in this assessment

Up to 12 euros.....

B. Allocators' Expectations of Whether Others Would Consider Transfer Fair [Self-Image Games]

For each possible transfer amount from 0 € to 12 € (in 1-€ steps), you indicate how confident you are in your assessment that others would consider this amount fair. You decide completely anonymously and unobserved. It is therefore about what you expect about whether others would consider the transfer fair if they knew about your transfer. Select exactly one of the following levels per amount:

100% = completely confident
50% = rather confident
0% = not confident

Please note that you must expand the further queries yourself by briefly clicking on "+Add row".

1. You transfer 0 euros to the Recipient. How confident are you that others, if they knew about the transfer, would consider this transfer fair?

I am ... [] in this assessment.

2. You transfer 1 euro to the Recipient. How confident are you that others, if they knew about the transfer, would consider this transfer fair?

I am ... [] in this assessment

3. You transfer 2 euros to the Recipient. How confident are you that others, if they knew about the transfer, would consider this transfer fair?

I am ... [] in this assessment
Up to 12 euros.....

C. Allocators' Own Fairness Perception: What is the Fairness Norm?

In your opinion, how many euros should you transfer to the Recipient for your decision to correspond to a moral understanding of fair behavior? Please first assume that you have not completed the work task [order randomized].

Please divide the 12 euros [in the control group: points] according to a moral understanding of fair behavior between yourself and the Recipient.
[] You [] Recipient
Now please assume that you have completed the work task.
[] You [] Recipient

D. Social Value Orientation

You will be randomly paired with another participant. In 32 rounds, you choose between two allocations, A and B, each of which allocates points to you and the other participant. You receive the points that the allocation you choose allocates to you; the other participant receives the points that the chosen allocation allocates to them. The points are summed over all 32 rounds. At the end of the experiment, you receive 0.2 cent per point, i.e. for 1000 points you are paid €2.

The 32 allocation decisions will now be displayed to you one after another. In each case, select your preferred allocation and then click "Continue".

Allocation Decision 1 of 32 (1 Point=€0.002)

Please select your preferred allocation:

A	B
For You 500 Points **For Your Partner** 0 Points	**For You** 304 Points **For Your Partner** 397 Points
Click here for Allocation A	**Click here for Allocation B**

The Following Table Shows all 32 Allocation Decisions:

	Alternative A		Alternative B	
	You	Partner	You	Partner
1	0	500	500	397
2	304	397	354	354
3	354	354	397	304
4	397	304	433	250
5	433	250	462	191
6	462	191	483	129
7	483	129	496	65
8	496	65	500	0
9	500	0	496	-65
10	496	-65	483	-129
11	483	-129	462	-191
12	462	-191	433	-250
13	433	-250	397	-304
14	397	-304	354	-354

	Alternative A		Alternative B	
	You	Partner	You	Partner
15	354	-354	304	-397
16	304	-397	0	-500
17	0	-500	-304	-397
18	-304	-397	-354	-354
19	-354	-354	-397	-304
20	-397	-304	-433	-250
21	-433	-250	-462	-191
22	-462	-191	-483	-129
23	-483	-129	-496	-65
24	-496	-65	-500	0
25	-500	0	-496	65
26	-496	65	-483	129
27	-483	129	-462	191
28	-462	191	-433	250
29	-433	250	-397	304
30	-397	304	-354	354
31	-354	354	-304	397
32	-304	397	0	500

E. Hypothetical Transfer without Work Option

Please imagine you had not received the opportunity to complete the work task. How would you have divided the 12 euros in this case?
You [] Recipient []

F. Hypothetical Transfer Decision Before Completing the Work Task

Please imagine you had first made your transfer decision and only afterwards were informed about the opportunity to complete the work task. Would you then have completed the task?
Yes [] No []

IV. Instructions for the Observer: Fairness Norm with and without Completed Work Task

Note: Before the impartial Observer makes their assessment, the Observer first receives the Allocator's instructions. In Treatment 2, they receive the complete instructions. In Treatment 1, however, the part of the instructions explaining to the Allocator that they can choose whether to complete the work task or not is not provided. If the Allocator has decided against completing the work task, the entire part of the instructions dealing with the work task is not provided. The Observer

therefore knows nothing about the work opportunity the Allocator had in this case. Then the instructions inform the Observer about their own task:

"Your task is to assess the fairness norm in two situations. First, you indicate what a fairly acting Allocator should transfer to the Recipient if the Allocator has completed the task. Then you give your assessment of what the Allocator should transfer if they have not completed the task. In both situations, you assess each possible transfer amount from 0 € to 12 € (in 1-€ steps) according to whether it corresponds to the fairness norm in your opinion and how confident you are in your assessment. Per amount, you select exactly one of the following levels. “Not confident” suggests that you do **not** view the transfer as fair.

100% = completely confident
50% = rather confident
0% = not confident

For this task you receive an endowment of €0.50, which you can use exclusively to reward the Allocator. If you clearly consider an amount to be fair, you award €0.50; if you are only rather confident, you award €0.25; if you do not consider the amount to be fair or are not confident, you award €0. The Allocator receives exclusively the amount you have awarded for exactly the transfer that the Allocator has actually made. Amounts not awarded from the €0.50 endowment you cannot keep and automatically flow back to the experiment management."