Tontrup, Stephan; Sprigman, Christopher Jon

Working Paper

Strategic Delegation of Moral Decisions to AI

Suggested Citation: Tontrup, Stephan; Sprigman, Christopher Jon (2025) : Strategic Delegation of Moral Decisions to AI, SSRN, Rochester, NY, https://doi.org/10.2139/ssrn.5696827

This Version is available at:
https://hdl.handle.net/10419/335206

Strategic Delegation of Moral Decisions to AI

Stephan Tontrup

Christopher Jon Sprigman[1]

Abstract

Our study examines how individuals perceive the moral agency of artificial intelligence (AI), and, specifically, whether individuals believe that by involving AI as their agent, they can offload to the AI some of their responsibility for a morally sensitive decision. Existing literature shows that people often delegate self-interested decisions to human agents to mitigate their moral responsibility for unethical outcomes. This research explores whether individuals will similarly delegate such decisions to AI to reduce moral costs. Our study shows that many individuals perceive the AI as capable of assuming moral responsibility. These individuals delegate to the AI and delegating leads them to act more assertively in their self-interest while experiencing lower moral costs.

Participants (hereinafter, "Allocators") took part in a dictator game, allocating a $10 endowment between themselves and a Recipient. In the experimental treatment, Allocators could involve ChatGPT in their allocation decision, at the cost of incurring added time to complete the experiment. When engaged, the AI executed the transfer by informing the Recipient of a necessary payment code. Around 35% of Allocators chose to involve the AI, despite the opportunity costs of a much-prolonged process.

To isolate the effect of the AI's perceived responsibility, a control condition replaced the AI with a non-agentive computer program, while maintaining identical decision protocols. This design controlled for factors such as social distance and substantive influence by the AI.

Allocators who involved the AI transferred significantly less money to the Recipient, suggesting that delegating the transfer to AI reduced the moral costs associated with self-interested decisions. This is supported by the fact that prosocial individuals, who face higher moral costs from violating a norm and thus would without delegation transfer more than proself individuals, were significantly more likely to involve the AI. A responsibility measure indicates that Allocators who attributed more responsibility for the transfer to the AI were also more likely to involve the AI.

---

[1] Stephan Tontrup is the Lawrence Jacobson Fellow of Law and Business, New York University School of Law. Christopher Jon Sprigman is the Murray and Kathleen Bring Professor of Law, and Co-Director, Engelberg Center on Innovation Law and Policy, New York University School of Law. We thank the NYU Stern School of Business and the Stern Center for Behavioral Research for allowing us to conduct the study in their laboratory. We thank Eric Mercadante for helping throughout the process, especially in organizing the laboratory sessions and coordinating the research assistants. We thank Anwar Ruff who helped us in using the subject pool of NYU`s CESS lab. We thank our research assistants, Sheikh Abubakar Asghar, Saanika Banga, Peichen (Heather) Li, Yeonju Kim, Jiarui Liu, Gayatri Menon Li, and Kasvi Vij, for their excellent and friendly support. We thank for their insights and comments Barton Beebe, Richard Brooks, Stefan Bechtold, Christoph Engel, Franco Ferrari, Jeanne Fromer, Talia Gillis, Yoan Hermstrüver, Johann Laux, Sunoo Park, Haggai Porat, Catherine Sharkey, Brian Sheppard, Holger Spamann, David Stein, Katherine Strandburg, Eyal Zamir, Tom Zur and many others and participants at workshops at the New York University School of Law.

The study suggests that AI systems provide human actors with an easily accessible, low-cost, and hard-to-monitor means of offloading personal moral responsibility, highlighting the need to consider in AI regulation not only the inherent risks of AI output, but also how AI's perceived moral agency can influence human behavior and ethical accountability in human-AI interaction.

I. Introduction

Current frameworks for "responsible" artificial intelligence systems are primarily concerned with the safety of AI's outputs rather than with AI's downstream effects on human behavior (Köbis et al., 2021). The European AI Act likewise centers on guarding against AI systems' inherent risks (Kaminski & Selbst, 2025). This article has a different focus. We ask whether interacting with AI will have a behavioral impact on AI users—specifically, we ask whether interacting with AI will make users more self-interested and more willing to violate norms.

AI and Delegation. As AI systems permeate individual and organizational decision-making, opportunities expand to collaborate with AI, and to delegate to it. Delegation is a common strategy for managing the psychological and moral costs of decision-making. It can shift who is or who appears to be responsible for a decision, in whole or in part, when the outcome is questionable or violates social or other norms. By sharing responsibility with an agent, or in some cases shifting it entirely, individuals may lessen anticipated blame, guilt, or reputational loss, which can encourage greater moral risk-taking (Hamman et al., 2010; Oexl & Grossman, 2013; Hill, 2015; Steffel et al., 2016; Bernstein et al., 2025).

This raises the question whether one of the reasons to delegate to AI will be to reduce or avoid moral responsibility for privately beneficial, yet unfair or otherwise normatively objectionable decisions. Whether that is true would depend on whether people consider AI to be a moral entity that may assume some or all of the responsibility and blame for an objectionable decision. If decision-makers indeed view AI as bearing moral responsibility, this would raise concerns that they can easily evade moral accountability by using a tool that is widely available, low-cost, and hard to monitor, creating a psychological "responsibility gap" that could encourage morally detached, irresponsible decisions (Cameron, 1999; Matthias, 2004; Carpenter et al., 2005; Simmler, 2024).

We analyze two mechanisms at play in delegation. First, if decisionmakers expect others to blame the agent, anticipated social-image costs—i.e., the harm to the decisionmaker's standing with others—decline. Moral psychology research shows that people sometimes ascribe moral responsibility to AI systems (Cameron, 1999; Engel, 2011; Gill, 2020; Stuart & Kneer, 2021; Feier et al., 2022; Leib et al., 2025). Second, if decisionmakers themselves believe agents bear a measure of moral responsibility, then inserting an intermediary may reduce self-image

costs—i.e., the harm to an individual's perception of their own moral character. Aside from these two mechanisms, delegation may also increase social distance, making it easier to violate other-regarding norms even when perceived responsibility is not reassigned and decision-makers and observers perceive AI merely as a tool (analyzing effects of social distance Charness & Gneezy, 2008; Hoffman et al., 1996; Frey & Bohnet, 1999). However, we assume that even if social distancing occurs, the opportunity to offload responsibility through delegation should afford AI users additional protection for both social-image and self-image versus social distancing alone.

Behavioral Self-Management (BSM). In recent years, we (along with our colleague Jennifer Arlen) have developed a line of research we term Behavioral Self-Management (BSM). BSM describes the strategic redesign of one's normative decision environment to mute the costs attending a particular decision—such as the self-image and social-image costs of violating fairness norms. BSM is not a single strategy but a family of strategies. Individuals may, for example, shift fairness norms toward a more favorable standard that places lower normative demands on the decision-maker. They may do so by exerting visible effort so that observers evaluate outcomes against a norm based on effort or merit (which tolerates more unequal outcomes) rather than a stricter equality norm (Tontrup, Arlen & Sprigman, 2025a). Or, they may strategically access information about others' compliance with an equality norm when compliance is expected to be low, thereby updating their beliefs so that noncompliance appears less morally costly, both to themselves and to others (Arlen, Sprigman & Tontrup, 2025b). Delegation, which we investigate here, is another BSM strategy.

In our prior work, we show that BSM also encompasses strategies outside the moral domain, such as sharing responsibility with an agent to reduce anticipated regret that prevents beneficial trading (Arlen & Tontrup, 2015a), selecting a group of professionals and following their choices to limit anticipated regret over a potentially bad decision (Arlen & Tontrup, 2015b), or by exploiting one's own loss aversion as a commitment device to improve performance (Tontrup & Sprigman, 2022). In particular, we showed that the individuals most likely to use these strategies were those with loss aversion and a strong propensity for regret.

In the context of fairness decisions, our BSM theory predicts that prosocial types are more likely to employ BSM. In the absence of BSM, higher self- and social-image costs lead prosocials to choose more other-regarding actions. Because they start from a higher baseline of other-regarding behavior, prosocials have greater scope than proselfs to curtail that behavior when a BSM opportunity becomes available, giving them stronger incentives to use BSM strategies. This mechanism is consistent with evidence that prosociality predicts norm compliance (Bandura, 1999; Fiedler et al., 2013; van Lange et al., 2013; Thielman et al., 2020), and that compliance is motivated by self-image and social-image concerns (Andreoni & Bernheim, 2009; Ariely et al., 2009; Matthey & Regner, 2015). And so, in our experimental setting here we predict that participants will be more likely to pursue a delegation-based BSM strategy the more prosocial they are and the more responsibility they attribute to the AI for the self-interested action they choose to delegate.

Experimental Design. We test these ideas in a laboratory study using an actual ChatGPT interface in one treatment and deterministic software (LimeSurvey) as the control. Allocators played two dictator games with a $10 endowment in each game: a Self-Image Game

(unobserved) and a Social-Image Game (with an Observer). In the AI Treatment, Allocators could decide directly what amount to transfer to the Recipient, or delegate part of the transfer process to ChatGPT. ChatGPT executed a scripted protocol that was explicitly explained to participants to prevent any inference that the AI made substantive recommendations about the "correct" amount to transfer to the Recipient; the explanation made clear that the protocol merely presented the transfer process. LimeSurvey implemented the same protocol.

Our design placed AI in a position analogous to a human agent performing the final action in a causal chain—an action typically attributed a large share of responsibility for outcomes even when the principal instructs and controls the agent (Slonim & Roth, 1998; Gerstenberg & Lagnado, 2012). In our study, the final action was carrying out the transfer. Subjects collected their earnings in envelopes marked with collection codes prepared before the experiment for subjects to pick up anonymously in a separate room and without further (human) assistance. Participants needed to know their code to receive payment; conveying that code was the last causal step required to effect the transfer. Our design assures that only the AI (or LimeSurvey) knew the code corresponding to each payment: Allocators did not know their code until revealed by the AI, and experimenters, while knowing the code list, did not know the amount a particular Allocator transferred and thus which code applied to any particular transfer. Accordingly, when participants delegated, the AI performed the final causal act by sending the code to the Recipient. In the LimeSurvey control, the messages were sent by the non-agentive LimeSurvey software.

By implementing strict decision protocols and holding them constant across treatments, our design allows us to rule out the possibility that treatment differences in delegation rates or allocation amounts are driven by reduced perceived responsibility from greater social distance, by perceptions that the AI may exert substantive influence, or by other framing effects of the transfer protocol.

We test two games in a strategy method design: participants complete the games consecutively and are randomly paid for one. In the Self-Image Game, participants' choices are not observed and when they decide to delegate, the reason must be to protect their self-image. In the Social-Image Game by contrast, we add a third party to the design, who observes the participants' decisions, thereby implicating the Allocator's social-image.

Key findings. Four central results emerge. First, delegation to AI was common: 38% of Allocators delegated in the Self-Image Game; 32% in the Social-Image Game. Participants preferred delegating to AI over delegating to non-agentive software (32% to AI vs. 10% to LimeSurvey in the Social-Image Game).

Second, transfers were lower in the AI Treatment than in the LimeSurvey Treatment, and, within the AI Treatment, delegators gave less than non-delegators. Because the LimeSurvey Treatment held constant whatever social distance was produced by participants not having to perform the transfer, the additional reduction in transfer amounts we observe in the AI Treatment is consistent with responsibility offloading to the AI being an effective BSM mechanism beyond mere social distancing.

Third, participants in both games attributed significantly more responsibility to AI than to LimeSurvey, and correspondingly less to themselves. Allocators were equally likely to delegate

and transfer similar amounts in both games—suggesting that Allocators believed AI could reduce their responsibility not only in their own view (preserving their self-image), but that impartial observers would also attribute responsibility to AI, such that delegation would also shield social-image. Importantly, this belief was objectively correct: Observers assigned similar responsibility to AI, enabling Allocators to reduce both self-image and social-image costs.

Fourth, the Allocators who chose to delegate aligned with the type our BSM theory predicts has a stronger behavioral reason and opportunity to delegate. Allocators who attributed more responsibility to AI had the opportunity to benefit from delegation and were also more likely to delegate (overall AI responsibility attribution among delegating Allocators was 27.6%, while participants who did not delegate attributed only 9.5%; participants attributed only 5.6% responsibility to the non-agentive LimeSurvey software). Moreover, as expected, attributing more responsibility to the AI predicted lower transfers, supporting the view that delegation allowed Allocators to offload responsibility rather than merely increase social distance to the Recipient. Additionally, those with stronger other-regarding preferences—that is, prosocials who face higher moral costs when violating distributive fairness norms and thus were motivated to delegate—were also more likely to delegate than proself types. Prosocial participants who also attributed some responsibility to the AI had a 52.4% likelihood of delegating (65.9% of delegators showed these characteristics). Consistent with our theory, subjects without other-regarding concerns did not delegate.

Contributions. First, our BSM study contributes to the literature on principal/agent relationships by showing that human-to-AI delegation follows a responsibility-shifting logic analogous to human-to-human delegation (Balliet et al., 2009; Hamman et al., 2010; Gawn & Innes, 2019; Gawn & Innes, 2021). Delegation to non-agentive software—which, in contrast, mainly creates social distance rather than shifting responsibility—occurs less frequently in our study and does not drive reductions in transfer amounts to the extent that delegation to AI does. This indicates that the effectiveness of delegation as a BSM strategy hinges on AI's perceived moral agency and the responsibility that AI agents are capable of bearing in both actors' and observers' eyes.

Second, our results inform the emerging debate about the moral agency of AI. While prior work (Cameron, 1999; Engel, 2011; Stuart & Kneer, 2021) shows that people are prepared to attribute responsibility to AI, we demonstrate that given the option and the right social context, people will exploit such attribution to offload responsibility for unethical decisions that advance their self-interest. And we show that such attribution is not simply self-serving but that impartial observers share that view and blame actors less. Higher perceived AI responsibility predicts both delegation and lower transfers among Allocators, showing that responsibility attribution is not merely a cognitive quirk but a lever for managing the self-image and social-image costs of normatively objectionable behavior.[2]

[2] Much of the algorithm–human interaction literature examines overreliance on automated advice and its harms (Dietvorst et al., 2015; Bigman & Gray et al., 2018; Logg, Minson & Moore, 2019). Our BSM account asks a different question: principals delegate to AI not because they "over-rely" on its computational capacity (Leib et al., 2025), but to obtain moral cover for an unfair transfer decision they want to make.

Third, our work has broad implications for law and policy. Most prominently, our study suggests that AI delegation can create space for unethical or irresponsible decisions—even among those who would otherwise comply with norms—under the socially accepted cover of AI mediation. For example, companies may strategically promote the perception of AI's moral agency to entice employees into unethical behavior, encouraging them to take greater risks and violate social or even legal norms. Our study suggests that offering a blend of incentives for achieving positive outcomes, while providing employees with the opportunity to offload responsibility onto AI, might motivate employees to leverage AI tools in pursuit of personal gain when doing so contradicts norms. This may be beneficial for the company, and provide an immediate payoff to the employee. Consequently, the rules of social and legal accountability can be undermined. And since prosocial types are particularly prone to offload responsibility—while proself types may have fewer concerns about taking normative risks at the expense of others from the outset—diminishing personal responsibility in those who would otherwise uphold norms may shift an organization's culture from an equilibrium of perceived widespread compliance with law and ethical standards to one in which noncompliance appears more common and is reinforced by social comparison (see Arlen, Sprigman & Tontrup, 2025c).

Our results also suggest that structuring the human-AI interaction can influence the moral agency that people attribute to AI and therefore also whether AI users can exploit this perceived moral agency. By manipulating whether AI appears to perform the final causal act in human-AI interaction, design choices shape responsibility attribution. For example, the AI may execute a decision by default unless the operator vetoes it. Although the operator retains control, this design allows the operator, by their own inaction, to let the AI make the final determination and offload responsibility to the AI.[3] AI could be deceptively framed as having final decision-making authority by presenting outputs as the AI's "decisions" when they are actually determined by human designers or deployers. This framing enables users to attribute moral agency and responsibility to the AI itself, thereby reducing the perceived moral cost of self-interested actions, compared to if the AI system was designed to disclose that a human principal directed its actions. In this way, an individual or firm deploying the AI—typically, an employer—may benefit from the unethical or unlawful behavior of an employee who uses the employer's AI without ever expressly instructing or encouraging that behavior.

On the other hand, clear disclosure of the AI's limited discretion could make it harder for users to offload responsibility to the AI system, thereby recentering accountability on the principal. These examples suggest that the empirical perception of moral agency is architectable: technical and procedural features can shift perceptions of normative accountability and thereby the behavior of those interacting with the AI.

However, since both companies and employees can have incentives to exploit the perceived moral agency of AI, business ethics measures alone—such as awareness campaigns or internal codes of conduct in dealing with AI (Hagendorff, 2020; Munn, 2023)—will likely not be effective, without further enforcement. Constraining strategic use of AI to offload accountability will require regulatory interventions. Custers et al. (2025) argue that distributing

[3] In this case felt responsibility would be further reduced by omission bias as people feel less responsible for outcomes they allowed by inaction than for outcomes they actively implement (Baron & Ritov, 2004).

legal responsibility across designers, deployers, and operators prevents legal "responsibility gaps" in AI environments and ensures more coherent accountability frameworks. Our results suggest that rules governing legal responsibility, when communicated to AI users, may also prevent psychological responsibility gaps. Mandating such disclosure may reduce moral wiggle room and the risk that deployers will entice users into ethically or legally dubious behavior. The U.S. Department of Justice has signaled that misuse of AI—especially when amplifying white-collar crime—may attract harsher penalties in the future, and that compliance programs must encompass AI risk mitigation. The European AI Act imposes stringent penalties—up to 7% of global annual revenue—for misuse or non-compliance, potentially signaling high stakes for organizations that intentionally structure AI to erode accountability. But as the general focus of AI regulation is on the inherent risks of AI outputs, it is far from certain whether this emerging regulation is prepared to capture the normative risk that actors in human-AI interaction may present when exploiting AI's perceived moral agency for their self-interest.

In Parts II and III of this Article, we describe our experimental design and methods. In Part IV, we detail our hypotheses and report our results. We close in Part V with a discussion of implications.

II. Experimental Design

A. Games and Treatments

We use the dictator game as the basic design of our study. The core game has two players, the Allocator and the Recipient. Roles are randomly assigned, and players are randomly matched. The Allocator is given an endowment of $10 and can decide how much of this endowment to keep or transfer to the Recipient. Each subject in the experiment will play both the Allocator and Recipient roles; however, subjects only learn of their role as Recipient when the experiment is finished and they receive the payment from an Allocator assigned to them. We employ two treatments in our study: the AI Treatment and the LimeSurvey Treatment.

1. AI Treatment

In the AI Treatment, participants complete the experiment directly in ChatGPT's chat interface—which, for brevity, we refer to hereafter as "the AI." We instructed the AI to strictly follow a pre-written script. If participants posed questions, the AI redirected them to our lab assistants rather than answering on its own. Without this restriction, free-form dialogue with the AI would have introduced uncontrolled variation across subjects and treatments, compromising experimental control. Yet, because our research question asks whether people perceive AI as capable of assuming moral responsibility, this conservative design likely works against our hypothesis—unrestricted conversation may have reinforced participants' tendency to attribute moral responsibility to the AI.

In the dictator game, Allocators could choose whether to involve the AI in their transfer decision or not. The AI offered to carry out the transfer for the participant, following a protocol which is explained in detail to the subjects before they decide whether to involve the AI. Control questions check that the Allocators have understood the protocol correctly.

According to the protocol, if the Allocator decides to involve the AI, the AI will go through a process to prompt the Allocator to specify how much to transfer. Initially, the AI will offer to transfer $0 and ask the Allocator if they agree or if it should transfer more. The Allocator must make an affirmative decision and either consent to or reject the amount. If the Allocator agrees to the amount, the AI will make the transfer. If the Allocator rejects the amount, the AI will offer a higher amount, moving first to $1. The process will continue until the Allocator agrees to a transfer, possibly reaching the maximum amount of $10. The protocol is structured this way to avoid suggesting to the Allocator that the AI has some information about or insight into what may be the "appropriate" allocation. If the Allocator decides not to delegate to the AI, they can directly specify the amount they want to transfer.

For our study's objective of analyzing whether Allocators will attribute responsibility to the AI, it is crucial that the experimental design ensures that enacting the transfer depends on an action of the AI. We made this action the final step in the causal chain that leads to implementing the transfer, as studies on the attribution of responsibility between multiple actors (Spellman, 1997; Gerstenberg & Lagnado, 2012) show that people tend to attribute significant responsibility for an outcome to the final actor in a causal chain of actions.

To this end, we prepared all possible transfer payments $0-10 in advance and provided them in envelopes marked with a numerical "collection code" specific to each transfer amount. We placed these envelopes in a separate room of the lab for participants to pick up. Participants could only pick up their payment if they knew the correct collection code. The instructions did not inform the Allocator of the codes; only the AI knew them in advance. Therefore, the Allocators could not execute the transfer by themselves without the AI informing them of the collection code associated with the amount the Allocator transferred.

When the Allocator chose to delegate the transfer, the AI communicated the collection code directly to the Recipient—this act was the last necessary for enacting the transfer; the Recipient was then able to claim their pre-prepared earnings without another person's further action.

Note that also the experimenters did not have the collection code that applied to any particular transfer, since we did not know how much the Allocator had transferred to the Recipient. We could only have learned the true amount by asking the participants or by checking their ChatGPT interface. However, this would have violated the experimental rules and required subjects' consent. This emphasizes the importance of the role that the experiment assigns to the AI for implementing the payments of the participants.

Consequently, following the cognitive theories of responsibility attribution referred to above, Allocators could be expected to attribute responsibility for the transfer to the AI if they indeed perceived it as capable of assuming responsibility, similar to how they might attribute responsibility to a human agent performing the last act in a causal chain.

By contrast, when the Allocator did not delegate, the AI provided the collection code only to the Allocator, who then had to inform the Recipient of the code themselves. Thus, when the Allocator did not delegate, it was the Allocator rather than the AI that performed the final action necessary for the transfer to come into effect, thereby diminishing the responsibility that the Allocator may attribute to the AI for its part in the transfer process.

Participants received two payments: first, in their role as Recipients for which they received the collection code, and second, in their role as Allocator, for which they received payment in a second envelope. This second envelope was also laid out in the separate room and marked with the participant's personal five-digit code that they had been asked to create and indicate online and on their paper-based questionnaire. As we could not learn the amount of any participant's payment from the ChatGPT interface without breaching their anonymity, participants indicated the amount they had transferred in a questionnaire they had to fill out at the end of the experiment. To verify their statement, they had to provide the collection code of the Recipient the AI had them informed of—a code that they could only have learned if they had actually transferred the particular amount they indicated.

As ChatGPT cannot currently send emails on its own, it was unable to inform the participants of their collection code via email. Instead, we used a Zoom connection to show participants the transfer message written by the AI in the Chat interface of the Allocator matched to them. We achieved this by leaving one booth free next to each lab participant. Once they had completed their role as Allocators in the experiment and had completed the questionnaires, we informed them that they were also Recipients of another Allocator's transfer. They were then asked to stand up and move to the empty booth next to them. The screen in this booth showed the Allocator's screen with whom they had been matched. If the Allocator had delegated the transfer to the AI, they saw the transfer message sent by the AI; otherwise, they saw the transfer message sent by the Allocator. This allowed the participants to learn the collection code. Once they had received the code, they left the lab to collect their payment.

2. LimeSurvey Treatment

In this second treatment, participants did not interact with ChatGPT; instead, the experiment was presented to them by LimeSurvey. While the name of the software and platform was not revealed, it was salient to participants that the survey software and the program it ran were deterministic and not an AI. Otherwise, we made sure that the interface looked similar to that of ChatGPT. In contrast to the AI Treatment, Allocators could not decide whether to delegate the transfer to the software; the software carried out the transfers of all participants in the LimeSurvey Treatment.

However, when Allocators had decided to delegate the transfer, the scripted decision-making protocol was exactly the same as in the AI Treatment. The program would first offer to transfer $0 if the Allocator agreed, and if the Allocator did not agree it would next offer to transfer $1 and so forth. When the Allocator affirmatively agreed to a transfer amount, the program sent the Recipient the collection code and the Recipient could pick up their transfer. If this protocol—by starting at $0—anchored delegators toward lower transfers, the same effect would arise under both the AI and the LimeSurvey treatments and thus cancel out.

Second, Allocators may also delegate to reduce their own responsibility for the Recipients' outcomes, even when they do not think that the AI is capable of bearing moral responsibility: delegation likely increases social distance between the Allocator and the Recipient, which should alleviate the sense of responsibility Allocators feel for final outcomes. The LimeSurvey Treatment controls for this social distance effect. By providing the essential collection code and writing the message to the Recipient, LimeSurvey relieves the Allocator of the moral burden of

the last action and may thereby increase the social distance between Allocator and Recipient in the same way as the AI does. The effect of social distance should thus be canceled out ceteris paribus when we compare the treatments.

Moreover, if participants attributed responsibility to the software programmer or to the experimenter who designed the decision protocol, such attributions would apply equally across treatments and thus cannot account for the observed treatment differences. This allows us to attribute any observed treatment difference in delegation and transfer choices to differences in perceived moral agency between the AI and LimeSurvey that allow Allocators to reduce responsibility for their transfers and lead Observers to view them as less accountable.

3. Within Comparison of Transfers and Delegations

We asked AI Treatment participants who had decided to delegate whether they would also have delegated to deterministic software. In the LimeSurvey Treatment, where the software carries out all transfers, we asked subjects whether they would delegate the transfer to LimeSurvey if that was an option, or to ChatGPT.

In the AI Treatment delegators also indicated what amount they would have transferred if they had not delegated to the AI. To incentivize this decision, we used a random-incentive mechanism: with probability 1 in 10, the stated transfer without delegation option was implemented in addition to payoffs from the two Self-Image and Social-Image Games. When implemented, the Allocator kept the untransferred amount, and the matched Recipient received the amount the Allocator stated they would transfer.

4. Social-Image and Self-Image Games

The subjects played two dictator games consecutively, the Self-Image Game and the Social-Image Game. The two games differ in that in the Social-Image Game an Observer was involved who would judge the Allocators' delegation and transfer decisions. The Observer in the Social-Image Game is given an endowment of $0.50 and asked to state what a fair Allocator should transfer to the Recipient and to reward the Allocator's transfer with the $0.50 if they perceive it as fair.

Although the Allocator knows the Observer will evaluate their decision, the prospect of an Observer-granted reward does not change the Allocator's financial incentives: keeping a dollar always yields a higher payoff than transferring an additional dollar to secure a potential $0.50 reward.

The Self-Image Game allows us to determine whether Allocators will involve the AI if their transfer decisions cannot be associated with them. We use a double-blind protocol where neither the experimenters can identify the participants nor do participants know the experimenters until the end of their session. The Allocators knew that the Recipients were not informed with whom they had been randomly matched and would receive the amount they transferred anonymously in an envelope. They also knew the Recipient would not learn whether they had delegated to the AI. If Allocators still chose to delegate in this setting, it suggests their decision is driven by self-image concerns—that is, that they delegate in order to justify their transfer decisions to themselves.

B. Questionnaires

Following the computer-based games, we asked participants to complete a number of measures on paper.

1. Responsibility Attribution

We asked Allocators the extent to which they would attribute responsibility for the transfer and its outcome to the AI, and the extent to which they would attribute it to themselves, and whether others would agree with their perception. Prior to the main study, we also queried participants acting as Observers how they would divide responsibility for the transfer between the AI and the Allocator subjects. Observers also decided whether to reward the $0.50 to the Allocators and made this decision for all possible allocation outcomes.

2. Norm Expectations

We elicit the Allocators' perception of what transfer-amount they think they should transfer, as well as what the Observers consider a fair amount to transfer. Allocators were queried after they made their transfer choice to ensure that the belief elicitation would not influence their transfer decisions (see Gächter et al., 2010).

3. The 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 their own. We used the ring measure developed by van Dijk, Sonnemans and van Winden (2002). The measure consists of 32 binary dictator games. Subjects were randomly matched with a partner and had to choose one of two allocations which would assign points to themselves and to their partner. The allocations vary in how much the partner gains from choosing the more generous option and how much the allocator must give up to provide that gain. The measure was incentivized yielding payoffs up to $4, depending on the subjects` own and their partners' choices.[4]

Social-value orientation has been shown to be predictive of cooperative behavior in many studies and across different games: for example, in public goods games studied in De Cremer and van Lange (2001) and in Fiedler et al. (2013), in one-shot and repeated prisoner's dilemma tasks in Balliet, Parks and Joireman (2009) and Pletzer et al. (2018), in investment games in Kanagaretnam et al. (2009), and in trust games in Yamagishi et al. (2013) and Haesevoets et al. (2014). The measure also proved reliable to predict for behavior in the field (for example van Lange et al., 2007). Hollander-Blumoff has analyzed consequences of social value orientations for doctrine across different legal fields (1997). For a detailed description of the measure see the Appendix.

The construction of the ring measure is important for understanding our hypotheses. The SVO is a continuum; most subjects are not purely selfish or even competitive, nor are they

[4] Subjects received $0.02 for each point they assigned to themselves and $0.02 for each point that was assigned to them as a Recipient. For example, a perfectly self-interested type with neither positive nor negative other-regarding preferences would always choose the allocation that maximizes their own points yielding 1,000 points and thus a payoff of $2 plus the amount their matched partner assigned to them.

purely prosocial. They may choose in one allocation decision to benefit their own interests if their 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.

When we refer to the common distinction between proself and prosocial types (SVO angle <+22.5°), many individuals classified as proself will also have other-regarding preferences of some degree, while individuals classified as prosocial (SVO angle ≥+22.5°) may prioritize self-interest in some situations, while nonetheless their other-regarding preferences are relatively stronger overall.choice.[5]

To avoid the arbitrariness of this distinction in what is truly a continuum of increasing prosociality, we use in regression analysis the continuous SVO angle to examine the influence of prosociality on delegation decisions and on transfers.

However, for some predictions—such that other-regarding participants have a stronger incentive to delegate to the AI—we use a distinction that is different from the common prosocial versus proself classification. We distinguish between participants with and without other-regarding preferences, operationalized as those with SVO angles greater than 0° versus those with SVO angles of 0° or less.

C. Control Questions

We asked our subjects a set of control questions to ensure their comprehension of the study's instructions. For example, they had to show that they understood that delegating would not limit the transfer choices they can make. And similarly, subjects had to demonstrate that they understood that they could not infer from the decision-making protocol that the AI offers any suggestion of what they should transfer. The complete list of control questions can be found in the Appendix.

III. Methods

A. Recruitment

We invited the participants for our study from the established subject pools of the behavioral center at Stern School of Business and the NYU Economics Department's CESS laboratory; all participants were registered prior to our study in one or the other laboratory. The participants were mostly current students at NYU, predominantly from the Stern School of Business, economics and liberal arts departments, and the engineering school. Approximately 20% of the subjects were externals with no current affiliation to the university. The lab sent participants a standardized invitation email with a link to schedule their participation at the lab (the NYU Stern Behavior Lab) hosting the study. The link became inactive once used. The invitation informed participants of the amount of time they would need to complete the

[5] Thus, the binary classification of prosociality, even though often used for its vividity, groups together subjects with no positive other-regarding preferences (SVO angle ≤ 0°) and those with positive but weaker other-regarding preferences (> 0° < 22.5°), categorizing them as proself and distinguishing them from individuals with stronger positive other-regarding preferences (SVO ≥ 22.5°), referred to as prosocials.

experiment, but did not reveal the objective of the study. Since most subjects were on campus or nearby, we did not pay a separate show up fee.

B. Anonymity

To ensure participants' anonymity throughout the experiment and payment process, they were not required to identify themselves when entering the lab. Instead, participants were asked to invent a five-digit code and record it in both the computer session and the questionnaire. This code enabled us to link the two data sources and attribute them to the same subject. Allocators provided the 5-digit code in the questionnaire, and their payment was placed in the payment room in an envelope marked with the code for them to pick up. In their role as Recipients, they also received their payment anonymously in a separate envelope marked with the collection code, which they had learned at the end of the experiment either from the AI or the Allocator. For example, the collection code 341 referred to a transfer amount of $4, which participants could pick up in the payment room.

C. Demographics

To preserve subjects' anonymity, we also did not elicit their demographics in the lab; since the sessions were relatively small, collecting demographic information on-site would sometimes have allowed us to identify which participant had made what choices. However, since subjects were registered for Stern's or the Economics Department's laboratory subject pool, we applied filters in the participant management software to ensure that the subject sample we sent invitations to was balanced in respect to gender and age composition; we also made sure to cover a broader list of undergraduate majors.

IV. Hypothesis and Results

Our BSM theory has four building blocks. Our theory first suggests that individuals (A) change their normative decision environment by choosing to delegate the transfer. That means we expect delegation to reduce the responsibility that subjects attribute to themselves for their transfers and that others attribute to them. Thus, by delegating they can pursue self-interest with lower self–image and social-image costs.

Second, our theory posits that Allocators will (B) delegate to benefit from reducing their transfers and (C) that they expect to manage their image concerns by reducing their perceived personal responsibility for an unequal outcome and the attribution by others of responsibility for that outcome. And finally, our BSM theory claims that Allocators act (D) behaviorally rationally: the more prosocial subjects are—i.e., the higher the moral costs associated with an unequal allocation—the more likely they should be to delegate. If we observe this behaviorally rational action, we can infer also that those who delegate can offload responsibility to the AI because they assume that AI can bear accountability and that others think so too.

A. Changing the Decision-Environment: Delegation

We predict that Allocators will delegate the transfer to the AI in order to reduce their perceived responsibility for the transfer decision. This requires that they and others perceive the

AI as an entity that can assume some share of the responsibility for an unequal allocation. We posit:

Hypothesis A_1: ***A significant number of Allocators of the type our theory predicts to have an incentive to delegate will delegate the transfer to the AI.***

In the Self-Image Game (n = 100), 38% of Allocators delegated; in the Social-Image Game (n = 100), 32% delegated. To assess whether a significant number of them did so in line with our theoretical expectations, we compare this rate to the proportion of participants who may have delegated because potentially they were confused by the instructions, or because of demand effects, or curiosity to interact with the AI.

Our theory identifies two conditions under which Allocators are more likely to delegate: (i) they exhibit other-regarding concerns (otherwise they have no need to delegate) and (ii) they believe delegation will shift responsibility to the agent, thereby reducing their self- and However, for some predictions—such as whether participants have an incentive to delegate to the AI—we use a distinction different from the common prosocial versus proself classification. Instead, we distinguish between participants with and without other-regarding preferences, operationalized as those with SVO angles greater than 0° versus those with SVO angles of 0° or less.

Because Allocators with a positive social value orientation ($\theta > 0°$) care about others, unfair transfers carry higher image costs for them. Delegating to an AI can reduce these costs by shifting responsibility, so we predict that the propensity to delegate rises with θ. Consistent with this theory, all of the subjects who delegate in either the Social-Image Game or the Self-Image Game exhibit positive other-regarding preferences in the social value orientation measure. Second, delegating subjects should attribute responsibility to the AI.

Both conditions—responsibility attribution and other-regarding concern—are met for 25/38 delegators (66%) in the Self-Image Game and 22/32 delegators (69%) in the Social-Image Game. However, some subjects (while exhibiting other-regarding concerns) delegated to the AI without attributing responsibility to it—n=10 in the Social-Image Game and n=13 in the Self-Image Game. These subjects may still engage in BSM and delegate to create social distance, making it psychologically easier for them to disregard the Recipient's outcome, but they cannot offload responsibility to the AI, which we expect to be more effective. However, they may also delegate for reasons inconsistent with BSM such as curiosity, demand effects, or confusion about the instructions.

To assess whether delegation is truly driven by this BSM strategy, we test whether delegation rates significantly exceed what could be explained by the theoretically inconsistent cases alone. To this end we conduct two binomial tests (one for each game) using 13% (Self-Image) and 10% (Social-Image) as benchmark rates. In both tests, we can reject the null hypothesis that the true delegation rate equals these benchmarks of (potentially) theory-inconsistent cases ($p < 0.001$). Thus, the tests suggest that we observe a significant proportion of delegation that aligns with our theoretical predictions.

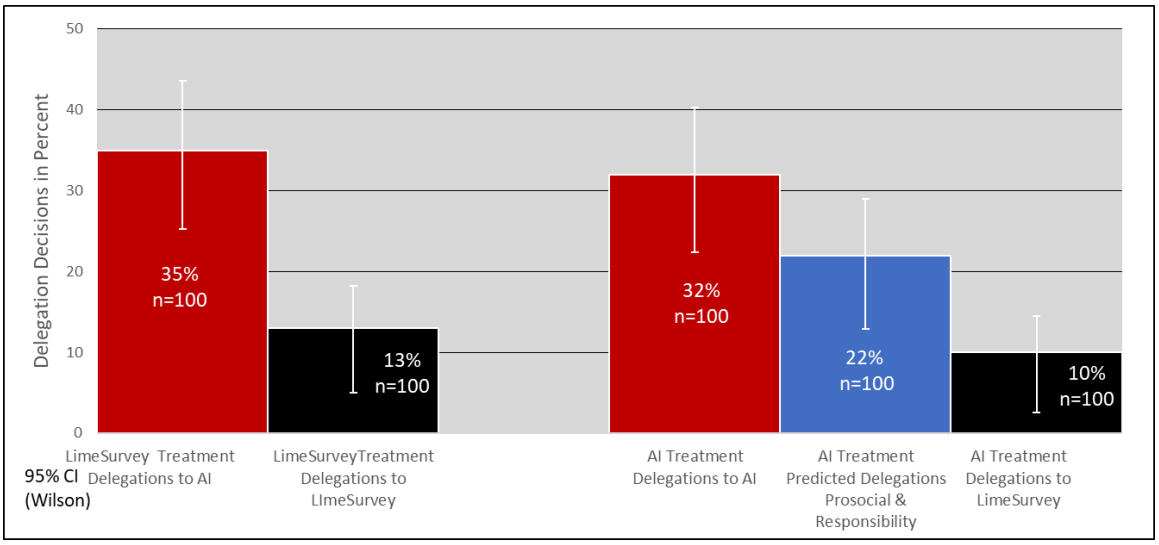
Second, we compare participants' delegation decisions to the AI with their hypothetical choice of delegating to deterministic survey software. Note that we collected the hypothetical

delegation choices only in the Social-Image Game. Based on the assumption that Allocators believe delegation to the AI can reduce perceived outcome responsibility, we hypothesize that participants will be more likely to delegate to the AI than to LimeSurvey, which we expect is unlikely to be viewed as meaningfully absorbing responsibility, as it appears to lack any autonomous agency. We posit:

Hypothesis A_2: ***Significantly more Allocators will delegate to the AI than to the non-agentive survey software.***

As we have seen, in the Social-Image Game of the AI Treatment, 32% of participants delegated their decision to the AI, however only 10% of the same participants indicated in the hypothetical questionnaire that they would have delegated the transfer to the non-agentive survey software. A McNemar test comparing these paired decisions yields $\chi^2(1) = 16.13$, $p< 0.001$, confirming that participants were significantly more likely to delegate to the AI than to the non-agentive software, as we hypothesized in A_2.

Figure 1: *Delegation Decisions across Treatments*



In the Social-Image Game of the LimeSurvey Treatment, participants were not offered a delegation option. Therefore, in the questionnaire we elicited two paired hypothetical choices: (i) whether they would have delegated the transfer to the (non-agentive) LimeSurvey software, and (ii) whether they would have delegated to the AI if these options had been available. While only 13% indicated they would have delegated to LimeSurvey if they were given the choice, 35% indicated they would have delegated to the AI. Comparing these paired responses yields a significant difference (McNemar $\chi^2(1) = 15.13$, $p < 0.001$). Figure 1 above depicts the results:

In **Table 1** we support the results with a logistic panel regression showing that in the Social-Image Game the odds for subjects delegating to the AI are significantly higher versus the same subjects delegating to LimeSurvey. In the AI Treatment, we compare subjects' actual delegation choices with their hypothetical choice whether to delegate to LimeSurvey. In a second panel regression we show the same difference in the LimeSurvey Treatment, comparing the two hypothetical delegation choices to AI and LimeSurvey that we prompted subjects to

make. Both regressions control for participants' social value orientation and the responsibility they attribute to the AI or to LimeSurvey.

Table 1: Delegation Decisions to AI vs. LimeSurvey in Social-Image Game

Variable Delegation Choices	**AI Treatment** Actual Delegation to AI vs. Counterfactual to LimeSurvey Logistic Panel (β, SE)	**LimeSurvey Treatment** Counterfactual Delegation to AI vs. Counterfactual to LimeSurvey Logistic Panel (β, SE)
Delegation option: to AI	0.787***	0.759***
(=1 to LimeSurvey=0)	(0.236)	(0.265)
Responsibility to AI	0.026***	0.009
(scale 0-100)	(0.008)	(0.007)
SVO (Angle)	0.015**	-0.007
	(0.006)	(0.005)
Constant	-2.406***	-1.454***
	(0.438)	(0.324)
Observations	200	200
Participants/clusters	100	100
Wald chi2(4)	19.60	9.58

Notes:
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Panel Logistic with Generalized estimating equations (GEE).
Robust standard errors clustered at subject level.
AI Treatment: Compares actual delegation to AI versus counterfactual delegation to LimeSurvey from same subject.
LimeSurvey Treatment: Both delegation preferences are counterfactuals.

The regression results show first, that participants have a strong preference for AI delegation, with coefficients of β= 0.787 and β= 0.759 in both treatments (p< 0.01), indicating that when given the choice, individuals are significantly more likely to delegate to AI than to LimeSurvey.

More importantly, responsibility attribution differs across delegation targets. In the AI Treatment, the more responsibility participants attribute to AI the more likely they are to choose AI delegation (β= 0.026, p< 0.01). In contrast, responsibility attribution shows no significant effect in the LimeSurvey treatment (β=0.009, p=0.320). This demonstrates that when people view AI as capable of bearing responsibility for outcomes, they become more willing to delegate to it the task of effectuating the transfer.

This pattern supports the hypothesis that responsibility attribution drives delegation decisions, but only when the delegation target—AI or LimeSurvey—is perceived as capable of bearing responsibility.

B. Gaining Benefits: Delegators Give Less than Allocators who Transfer Themselves

Having shown that Allocators delegate to alter their decision environment, we now examine whether, conditional on delegation, Allocators obtain a benefit by reducing their transfers. We posit:

Hypothesis B_1: ***Allocators in the AI Treatment transfer significantly less than those in the LimeSurvey condition.***

In both the Social-Image Game and the Self-Image Game, average transfers were lower in the AI Treatment than in LimeSurvey, consistent with B_1. In the Social-Image Game of the LimeSurvey Treatment, Allocators transferred \$2.83 on average, whereas Allocators of the AI Treatment transferred \$2.20 in the Social-Image Game; the treatment difference is statistically significant at t(198)=1.97, p=0.04. In the Self-Image Game of the LimeSurvey Treatment the mean was \$3.60 compared to \$2.74 in the AI Treatment, at t(198) =2.77, $p < 0.01$; this difference is also significant. We summarize the data in the table below.

Table 2: Descriptive Transfers across Games and Treatments

Game	AI Treatment Mean Transfer (n=100)	Lime Treatment Mean Transfer (n=100)	Transfer Gap (LS−AI)	t(df)	p-value
Social-Image	\$2.74	\$3.60	0.86	2.77 (198)	$p < 0.01$
Self-Image	\$2.20	\$2.83	0.63	1.97 (198)	$p = 0.04$

These Treatment differences in transfers hold in regression analysis. The Tobit models account for censoring at the bounds of the transfer distribution (\$0 and \$10) and with robust SEs clustered by subject. We code the treatment indicator as "AI" =1, and with "LimeSurvey"=0 as the reference category. The AI Treatment indicator is negative and statistically significant in both games, supporting that transfers are lower under AI. We report full regression results in **Table 3** below.

Table 3: Transfers by Treatments

Variable Transfer	**Social-Image Game** Tobit (β, SE)	**Self-Image Game** Tobit (β, SE)
AI Treatment (1=AI, 0=LimeSurvey)	-1.129*** (0.383)	-0.994*** (0.432)
Constant	3.528*** (0.237)	2.653*** (0.330)
Censored left	33	54
Censored right	5	2
Observations	200	200
Pseudo R^2	0.02	0.01
Wald F(1,199)	8.68***	5.02***

Notes:
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
Tobit censored at \$0 and \$10.
Robust standard errors clustered at subject level.

Note that we find significant treatment effects in transfers in both games despite the fact that social distance can affect decisions only in the AI Treatment in instances where Allocators choose to delegate. By contrast, in the LimeSurvey Treatment, where the software automatically enacts all transfers, social distance could influence the decision of all Allocators. LimeSurvey always informs the Recipient of the collection code whereas in the AI Treatment AI writes the message only for those Allocators who choose to delegate. And this means that a social distance effect, if present, is working against the treatment effect we report, and therefore we are likely underestimating the true treatment difference. We nevertheless designed this

treatment comparison because it creates a clean benchmark: it shows that the effect of the AI Treatment must go beyond inducing social distance and isolates the effect offloading responsibility to the AI has on transfers.

To further investigate the true impact of responsibility offloading, we limit the sample of the AI Treatment to delegators and compare their transfers with the transfers of the full sample of LimeSurvey subjects. We find that transfers of delegators in the AI Treatment are substantially lower in both the Social-Image Game ($1.96) and the Self-Image Game ($1.28), compared to the LimeSurvey transfers ($3.60 and $2.83 respectively).

Table 4: Delegator Transfers by Games and Treatment

Game	AI Treatment Only Delegators Mean Transfer (*n*=32; 38)	LimeSurvey Treatment Mean Transfer (*n*=100)	Transfer Gap (LS−AI)	t(df)	*p*-value
Social-Image	$1.96	$3.60	$1.64	3.94 (130)	$p < 0.0001$
Self-Image	$1.28	$2.83	$1.55	3.61 (136)	$p < 0.0001$

To show that these substantial differences are not driven by possible selection effects—i.e. that delegators who decide to involve the AI might be systematically less generous than the rest of the sample of the AI Treatment who did not involve the AI—we compare the transfers delegators made when not presented with a delegation option to the transfers those Allocators made who did not delegate. Note that we elicited a no-delegation transfer only in the Self-Image Game, and it was incentivized: with a 1-in-10 probability, the stated transfer was implemented. If implemented, the Allocator kept the untransferred remainder of the endowment and the Recipient received the amount sent; these payments were in addition to earnings from the main game.

We find that delegators (rather than being less generous) transferred $3.81 when they had no option to delegate—descriptively more, not less, generous than the non-delegators in the AI Treatment (+$0.21; t(98)=-0.561, p= 0.57).

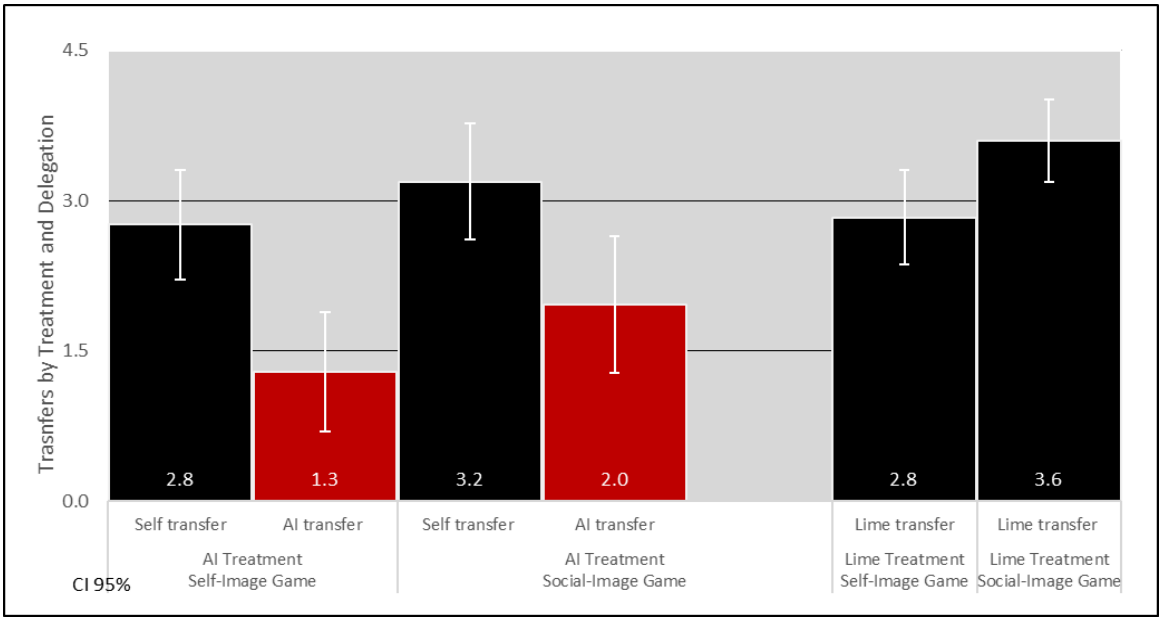
That delegators in the Social-Image Game are not a selection of a less generous part of the sample is further demonstrated by comparing the mean prosociality of the group of delegators (39.2 degrees) with the Allocators who did not delegate in the AI Treatment, who turn out to be significantly less prosocial (15.6 degrees; t(98)=-2.565; $p < 0.01$), in line with our theory that prosocials have a stronger incentive to delegate. **Figure 3** below in Section C. presents Allocators' responsibility attribution across treatments and game types.

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

Hypothesis B_2: ***Allocators who delegate the transfer to the AI will transfer significantly less than those who decide not to delegate.***

We find this hypothesis supported: In the Social-Image Game of the AI Treatment, delegating Allocators transferred $1.96 (n=32) on average, significantly less than the $3.10 (n=68) that non-delegating Allocators transfer (t(98) =2.341, $p= 0.02$), a difference of 36.8%.

Figure 2: *Transfers by Treatments and Delegation in the Social-Image & Self-Image Game*



The Self-Image Game shows likewise: delegators transferred a mean amount of $1.28 (n=38), whereas non-delegating Allocators averaged $2.75 (n=62; t(98) =3.523, $p< 0.001$), which yields a gap of 53.5%, suggesting that delegating leads Allocators to significantly reduce their transfers. The average transfers across treatments and games and conditioned on delegation are illustrated in **Figure 2** above.

We confirm these results in Tobit regression models that account for censoring at $0 and $10 and control for subjects' prosociality. We report the regression results in **Table 5**.

Table 5: Transfers by Delegation

Variable	**Social-Image Game**	**Self-Image Game**
Transfer	Tobit (β, SE)	Tobit (β, SE)
Delegation	-1.411**	-2.157***
(=1 if delegated)	(0.628)	(0.628)
Constant	2.383***	2.345***
	(0.726)	(0.394)
Censored left	25	35
Censored right	2	0
Observations	100	100
Pseudo R^2	0.02	0.03
Wald F(1, 99)	5.04	11.69

Notes:
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$
Analysis restricted to AI Treatment condition only.
Tobit censored at $0 and $10.
Robust standard errors clustered at subject level.

The delegation indicator (1= delegated, 0= made transfer alone) is negative and significant both in the *Social-Image Game* (β= −1.513, p= 0.02), and in the *Self-Image Game* (β= −2.187, p< 0.01).

Additionally, we asked delegating Allocators in the Social-Image Game how much they would have transferred had they not had a delegation option (recall that these responses were incentivized). Seventy-five percent—24 of 32—reported they would have transferred more.

When we replace delegators' transfers with these no-option-to-delegate transfer amounts (keeping the transfers of Allocators who did not delegate), the AI Treatment's hypothetical mean transfer rises to $3.33, nearly closing the gap to the LimeSurvey Treatment's mean of $3.60; the remaining difference of $0.27 is not statistically significant ($t(198) = 0.921$, $p = 0.358$).

This finding suggests that if subjects did not have an option to delegate their transfer decision in the AI Treatment, their transfers would have closely resembled those of subjects in the LimeSurvey Treatment, further supporting that the treatment difference is driven by delegation to the AI, as we hypothesized.

C. Protect Self-Image and Social-Image: Delegation Reduces Personal Responsibility for (Unfair) Transfers

We have seen support for the first two elements of our theory: **(A)** subjects change their normative decision environment by delegating to the AI, and **(B)** they do so for personal gain by reducing their transfers. We now turn to the third building block: **(C)** Allocators can use delegation to shield both their self-image and their social-image by offloading responsibility to the AI. For this mechanism to operate, Allocators must first perceive their responsibility to be lower when they delegate to the AI, and they must anticipate that Observers will hold them less responsible when they delegate. Finally—and just as importantly—Observers must actually attribute less responsibility to them.

First, we focus on Allocators and analyze whether (i) Allocators attribute responsibility to the AI and (ii) expect Observers to attribute less responsibility to them when they delegate. Second, we examine whether Observers actually do attribute less responsibility to Allocators who delegate.

1. Allocators: Perceived Own Responsibility for Transfers and Expected Responsibility Attribution by Observers

We elicited Allocators' self-attribution of responsibility for the transfer and compare it across treatments and contingent on delegation.

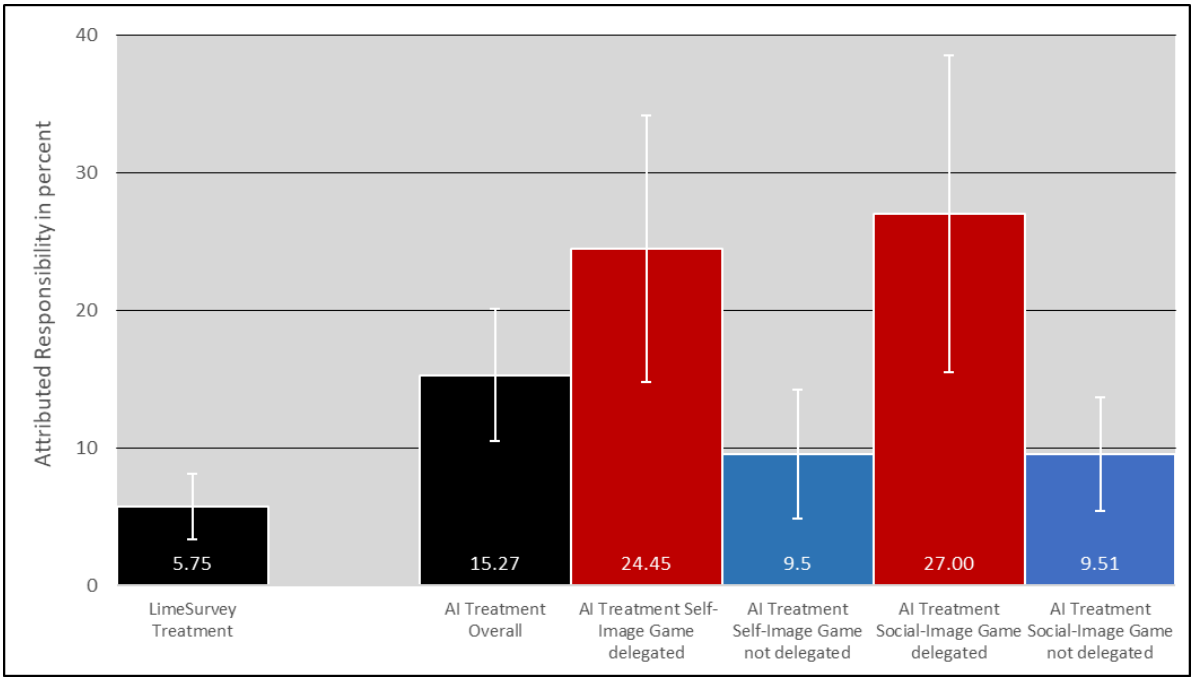
Hypothesis C_1: ***Allocators in the AI Treatment attribute more responsibility for the transfer to the AI than Allocators in the LimeSurvey Treatment.***

Hypothesis C_2: ***Allocators who delegate attribute more responsibility to the AI than non-delegators.***

Our results support both hypotheses. First, the level of responsibility attributed to the system is higher in the AI Treatment than in the LimeSurvey Treatment: Allocators assigned 15.27 points (0–100 scale; n=100) to the AI in the AI Treatment versus 5.75 points (n=100) to the deterministic software in the LimeSurvey Treatment; this difference is statistically significant ($t(198) = 3.496$, $p = 0.001$) supporting C_1.

Second, for testing C_2 we restrict the sample to Allocators who delegated in the AI Treatment. We find that the difference between treatments becomes more pronounced. In the Self-Image Game, delegators assigned 24.45 points (n=38) of responsibility to the AI, compared to the full sample of subjects in the LimeSurvey Treatment who assign 5.75 points (n=100) to LimeSurvey ($t(136) = 5.31$, $p< 0.001$). In the Social-Image Game, delegators assigned 27.0 points (n=32) to the AI versus the 5.75 points (n=100) the full sample of Allocators average in the LimeSurvey Treatment ($t(130)=5.76$, $p< 0.001$). The results for both games support C_2: Delegators attribute more responsibility to AI than non-delegators. We illustrate these results in **Figure 3** below.[6]

Figure 3. *Responsibility Attribution by Treatments and Delegation in the Social- & Self-Image Game*



Third, Allocators can only reduce their social image costs through delegation if they expect Observers to attribute responsibility to the AI and thus less to them.[7] We posit:

Hypothesis C_3: ***Delegating Allocators expect Observers to attribute less responsibility to them.***

Of 32 delegators in the Social-Image Game, 25 (78.1%) expected Observers to attribute responsibility to the AI and accordingly less to them when delegating, while only 7 (21.9%) did not. As in Section A, we treat this 21.9% group as a benchmark rate for delegation that is

[6] Note that while subjects were asked to attribute responsibility to the AI only once, the averages differ across Social-Image and Self-Image Games because the number of delegators differs between the two games (Self-Image: n=38; Social-Image: n=32). By contrast, in the LimeSurvey Treatment, where LimeSurvey carries out all transfers and the n is thus constant across games, the attributed responsibility value remains the same.

[7] We asked the same Observers for the responsibility value they would attribute to the AI and to LimeSurvey.

potentially inconsistent with BSM theory potentially driven by curiosity, misconception or demand effects. A binomial test rejects the null hypothesis that the true rate of delegation driven by the expectation that Observers will attribute lower responsibility equals this 21.9% benchmark ($p< 0.001$), in support of hypothesis C_3.

By contrast, of the 68 non-delegators of the AI treatment, 32 (47.1%) expected Observers to attribute responsibility to the AI, a 31.1% lower rate compared to delegators (Fisher's exact p = 0.0046). The result aligns with our theory that delegation is chosen to protect social image, supporting our BSM interpretation.[8]

2. Observers: Responsibility Attributed by Observers is lower in AI Treatment

Finally, we analyze Observers' attribution of responsibility. We posit:

Hypothesis C_4: ***Observers attribute less responsibility to Allocators for a transfer in the AI Treatment when delegated than when not.***

67.5% of Observers attributed responsibility for the transfer to the AI (27/40; exact binomial test against 50%, $p= 0.04$). Among those who attributed some responsibility to the AI (n = 27), almost all (26/27) stated that an Allocator who delegates to the AI is less responsible for the outcome than an Allocator who executes the transfer herself (exact binomial test against 50%, $p< 0.001$). This indicates that delegation to AI indeed provides Allocators with protection for their social-image.

When we compare the responsibility that Allocators attribute to the AI (n=100; 15.27 points) with the responsibility that Observers attribute to the AI (n=40; 18.4 points), we find little difference (t(139)= -0.629, $p= 0.53$) with Observers not falling short of Allocators' responsibility attribution to AI, suggesting that Allocators responsibility attribution may not subject to severe self-serving bias. This supports our theory that Allocators truly want to protect their social.image which they can only do if Observers in fact attribute less responsibility to them when they delegate.[9]

Finally, we find: when we compare the responsibility that Observers attribute to AI (18.4 points) compared to the non-agentive LimeSurvey software (4.12 points). A paired t-test yields that this difference is (n = 40; t(39) = 3.332, $p= 0.002$).[10]

---

[8] As we will show later, the rate of prosocials among the delegators is significantly higher than among non-delegators, which explains why non-delegators even though some expect Observers to attribute less responsibility to them when they delegate to the AI, nevertheless do not delegate. With low other-regarding concern their incentive is minor.

[9] Notice that subjects who delegated and transferred a low amount may have increased the responsibility they attributed to the AI in order to avoid cognitive dissonance between the transfer they made and their other-regarding preferences. This comparison to the Observers' attributions suggests that responsibility attribution was not driven by cognitive dissonance.

[10] Our sample includes 15 subjects from a pilot study we conducted initially. They were presented with identical materials as later participants in the role of observers and their choices do not differ from those of the participants in the main study in a measurable way. However, these subjects later completed the experiment serving as Allocators. While they were not informed they would take on this role, they may have anticipated this transition, potentially influencing their decision-making.

D. Strategic Behavior: Allocators' Delegation and Transfer Choices are Behaviorally Rational

The fourth element of our theory states that **(D)** individuals act behaviorally rationally when reducing their moral costs by delegating in order to limit their own responsibility. If this assumption holds, we should observe that the greater the self-image and social-image costs Allocators can avoid by delegating, the more likely they should be to delegate.

1. Delegation Choices

We assume that individuals' self-image and social-image costs from making transfers that violate the applicable fairness norm increase with their prosociality, since transfers in dictator games are highly correlated with individuals' prosociality. Thus, we expect prosocial Allocators to transfer more (absent delegation), providing prosocials with a stronger incentive to delegate.

Second, Allocators' decision to delegate should depend on how much responsibility they attribute to the AI; the more responsibility they attribute, the more likely they should be to delegate the transfer to the AI. Consequently, even equally prosocial Allocators may differ substantially in their incentives to delegate: they should only opt for delegation when they expect that shifting responsibility to the AI will sufficiently reduce both their self-attributed and externally-perceived responsibility, such that the social-image and self-image costs of breaching the fairness norm are outweighed by the financial gain from keeping a larger share.[11]

Our theory that Allocators make strategic, behaviorally-rational delegation decisions leads us to two hypotheses.[12] We posit:

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

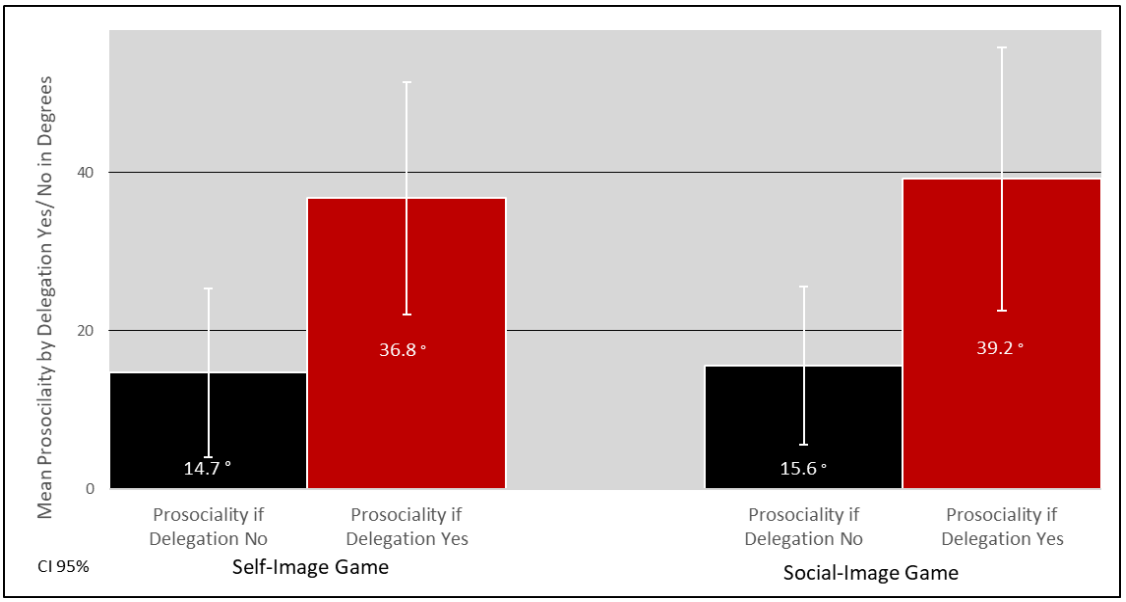
Hypothesis D_2 & D_3: ***The more responsibility for the transfer Allocators attribute to the AI and expect others to attribute to the AI, the more likely they are to delegate.***

In both games, we find significant support for D_1 and D_2. The average value of prosociality across all participants in the AI Treatment is 23.13 degrees. Among delegators in the Self-Image Game, we measured 36.77 degrees, while prosociality among those subjects who did not delegate is significantly lower at 14.77 degrees. ($t(98)=-2.4833$; $p= 0.01$). In the Social-Image Game it is 39.18 degrees among delegators and 15.57 for non-delegators ($t(98)= -2.565$; $p= 0.01$). We illustrate the results in **Figure 4** below.

[11] The measure distinguishes four social types; we indicate their prevalence in our sample in brackets: altruistic type (n=11), cooperative type (n=130), self-interested type (n=152) and competitive type (n=6). The distribution of types between the treatments is not significantly different.

[12] Transfers in dictator games (van Lange, 1999), in trust games (Bochet et al. 2006; van den Bos et al., 2009; Fetchenhauer & Dunning, 2012), and in public goods games (Balliet et al., 2009; Bogaert et al., 2008; Fiedler et al., 2013), have been shown to be highly correlated with individuals' prosociality.

Figure 4. *Prosociality by Delegation*



The average attribution of responsibility among delegators in the Social-Image Game is 27.0 points versus 9.51 among non-delegators ($t(98)= -3.526$, $p< 0.01$). In the Self-Image Game we observe 24.55 points among delegators and 9.64 points among non-delegators ($t(98)= -3.084$, $p< 0.01$). The regression analysis shows that an increase in responsibility attribution of one percentage point leads to an increase in the probability of delegation of 0.8 percentage points ($\beta= 0.008$). Accordingly, attributing responsibility to the AI made Allocators in the Social-Image Game 21.6 percentage points more likely to delegate, while in the Self-Image Game the attribution led to an increase of 19.6 percentage points, supporting our hypothesis D_2.

The data further support our theory that prosociality and responsibility attribution drive delegation decisions. Consistent with our theory, all the subjects who delegate in either the Social-Image Game or the Self-Image Game express positive other-regarding concern in the social value orientation measure (SVO>0°), while none of the subjects who lack other-regarding preferences decided to delegate. Of the delegators in the Self-Image Game, 65.7% (22/32), and in the Social-Image Game, 68.7% (25/38), also attribute responsibility to the artificial agent, supporting our hypothesis that prosocial individuals use the AI to offload responsibility to mute their other-regarding concerns. Those delegators who delegated to the AI without attributing responsibility to it, 34.2% (13/38) in the Self-Image Game and 31.25% (10/32) in the Social-Image Game, we assume still engage in BSM; they may delegate to create social distance, thereby making it psychologically easier for them to pursue self-interest. Finally, subjects who think AI bears responsibility for the transfer do nevertheless not delegate to the AI, if they exhibit no other-regarding concerns. This aligns with our theory, as they should have no reason to delegate.

For responsibility attribution, we find a stronger effect: the increase of one percentage point leads to an increase in the probability of delegation of 0.8 points (β=0.008). Accordingly, the average attribution of responsibility among delegators in the Social-Image Game (27.6 points) made these subjects 22.13% more likely to delegate, while in the Self-Image Game the

attribution of 24.55 points led to an increase of 17.19%. The average attribution of responsibility across all Allocators including those who did not delegate is 15.27 points leading to an increase in delegation probability of 12.22% in the Social-Image Game and 10.69% in the Self-Image Game. Non-delegators attribute only 5.7% to AI.

Consistent with our theory, all the subjects who in fact delegate in either the Social-Image Game or the Self-Image Game express positive other-regarding concern in the social value orientation measure (SVO>0°). However, some subjects delegated to the AI without attributing responsibility to it: 10 in the Social-Image Game and 13 in the Self-Image Game. Note that these subjects might still engage in BSM; they may delegate to the AI to create social distance, making it psychologically easier for them to pursue self-interest and suppress other-regarding concerns.[13] The table below shows that among delegators, 65.7% exhibit other-regarding preferences and attribute responsibility to the AI, supporting our hypothesis that prosocial individuals use the AI to offload their felt responsibility over the unfair transfer. Second, 34.2% show other-regarding concern but do not attribute responsibility to the AI, consistent with delegating to the AI in order to increase social distance between themselves and the recipient. Finally, also supporting our theory, subjects that exhibit no other-regarding concerns and therefore have according to our BSM theory no reason to engage in BSM also do not delegate to the AI.

Table 6: Other-Regarding Concern & Responsibility Attribution Predict Delegation

Theory	Predictors	Self-Image Game Delegation Rate	Social-Image Game Delegation Rate
Responsibility offloading	Other-regarding concern AND responsibility attribution	Cond. on delegation 65.79% In total sample 25%	Cond. on delegation 68.75% In total sample 22%
Social Distance	Other-regarding concern BUT NO responsibility attribution	Cond. on delegation 34.21%	Cond. on delegation 31.25%
No Reason to Delegate	NO Other-regarding concern	Cond. on delegation 0%	Cond. on delegation 0%

Notes:
"Conditional on delegation" means the percentage is calculated only among participants who delegated.
"Total sample" means the percentage is calculated out of all participants in that treatment.

To confirm our findings, we estimate logistic and linear probability models for the Social-Image and Self-Image Games, with delegation as the dependent variable and prosociality as well as responsibility attribution as predictors. In both Games, we find significant support for D_1 and D_2. In the Social-Image Game, prosociality (logistic: OR = 1.012, p= 0.04; LPM: β=

[13] As shown in Section A, the proportion of subjects who delegate, have other-regarding preferences and attribute responsibility to the AI is significant in the total sample when tested against the benchmark of subjects who delegate without attributing responsibility to the AI ($p < 0.001$).

1.012, p= 0.04) and responsibility attribution (logistic: OR = 1.027, p< 0.001; LPM: β= 1.012, p= 0.04) significantly predict delegation. Holding responsibility attribution constant, the more prosocial Allocators are more likely to delegate.[14] Similarly, holding prosociality constant, higher responsibility attribution to the AI increases the probability of delegation. In the Self-Image Game, the same pattern emerges: prosociality (OR = 1.016, p< 0.01; LPM: β= 0.002, p< 0.01) and responsibility attribution (OR = *1.048, p< 0.01;* LPM: *β= 0.008, p< 0.01)* both significantly increase the odds and probability of delegation.[15] We report these results in **Table 6**.

Table 6: Delegation Decisions by Responsibility & Prosociality

Variable	Social-Image Game		Self-Image Game	
Delegation	**Logistic OR (SE)**	**LPM β (SE)**	**Logistic OR (SE)**	**LPM β (SE)**
SVO (Angle)	1.016***	0.002**	1.014***	0.002***
	(0.006)	(0.001)	(0.005)	(0.001)
Responsibility	1.048***	0.008***	1.041***	0.008***
(scale 0-100)	(0.015)	(0.002)	(0.016)	(0.002)
Constant	0.142***	0.138***	0.230***	0.211***
	(0.057)	(0.047)	(0.083)	(0.053)
Observations	100	100	100	100
Pseudo R^2 / R^2	0.22	0.253	0.170	0.200
Wald chi2(2)/ F(2, 99)	13.21	18.13	11.54	13.20

Notes:
* p < 0.10, ** p < 0.05, *** p < 0.01.
Robust SEs clustered at subject level.
LPM = Linear Probability Model; OR = Odds Ratio.
LPM coefficients represent percentage point changes in delegation probability.
Logistic regression odds ratios (OR) confirm consistent effects across specifications.

Second, we analyze whether Allocators' expectation that Observers will attribute responsibility to the AI and thus less to them, significantly predicts their delegation decision in the Social-Image Game. Note that the expectation is a binary variable, either Allocators expect Observers to attribute responsibility to the AI or they expect not to do it. As hypothesized in D_3, expected responsibility attribution by Observers significantly predicts delegation of Allocators (logistic: OR= 3.951, p< 0.001; LPM: β= 0.289, p< 0.001).

[14] For interpreting the effect of prosociality, the LPM model suggests that an increase of one degree in the ring measure leads to an increase in the probability of delegation by about a quarter of a percentage point (β=0.0025) in both games. Compared with a proself type with no other-regarding concerns at 0 degrees on the ring measure, moving to a moderately prosocial-cooperative type at 45 degrees yields a 12.1% increase in the probability of delegation. Coding prosociality as a binary variable reveals a 10.4% increase in delegation rates when moving from proself to prosocial types in both games.

[15] For interpreting the effect of responsibility attribution, the LPM model suggests that an increase of one point of attributed responsibility leads to an increase in the probability of delegation by about 0.8 of a percentage point (β=0.008) in both games. Compared with a subject who does not attribute responsibility to the AI, the average score of delegators in both games of about 20.1% increase in the probability of delegation and for non-delegators of 7.6%.

Table 7: Delegation Decisions by Expectation & Prosociality

Dependent	Social-Image Game	
Var	Logistic Reg	LPM
Delegation	OR (SE)	LPM β (SE)
SVO	1.013**	0.002**
	(0.006)	(0.001)
Responsibility	3.955***	0.289***
	(1.865)	(0.092)
Constant	0.187***	0.164**
	(0.079)	(0.071)
Observations	100	100
R^2	0.121	0.145

Notes

∗ p < 0.10, ∗∗ p < 0.05, ∗∗∗ p < 0.01.

Robust SEs clustered at the subject level.

LPM = Linear Probability Model; OR = Odds Ratio. LPM coefficients represent percentage-point changes in delegation probability. Logistic regression columns report odds ratios (OR) and confirm consistent effects across specifications.

To support that the effect of delegation hinges on participants offloading responsibility to the AI, rather than *only* on increasing social distance, we re-estimate the delegation model considering solely those participants our theory predicts have reason to delegate to reduce their responsibility—subjects who displayed both other-regarding concern and attributed responsibility to the AI. This is true for 42 subjects.

We test whether responsibility and prosociality significantly predict delegation in this sample. We estimate a linear probability model (logistic regressions yield the same results) and results find that both responsibility attribution and SVO are significant predictors of delegation. We present the results in **Table 8** below.

Table 8: Only Subjects with Other-Regarding Preferences who Attribute Responsibility

Variables	**Social-Image Game**		**Self-Image Game**	
Delegation	**Logistic OR (SE)**	**LPM *β* (SE))**	**Logistic OR (SE)**	**LPM *β* (SE)**
SVO (Angle)	1.023**	0.004**	1.032**	0.004**
	(0.011)	(0.001)	(0.013)	(0.001)
Responsibility	1.051**	0.008**	1.039**	0.006***
(scale0-100)	(0.021)	(0.002)	(0.016)	(0.02)
Constant	0.093***	0.098*	0.135**	0.179
	(0.068)	(0.100)	(0.110)	(0.119)
Observations	42	42	42	42
Adj/ Pseudo R^2	0.25	0.24	0.21	0.24
Wald chi2(2)/ F(2, 41)	8.28	10.96	7.47	6.50

Notes:

∗ p < 0.10, ∗∗ p < 0.05, ∗∗∗ p < 0.01.

Only participants who our theory suggests have reason to delegate: angle >0 & responsibility attribution >0;

LPM = Linear Probability Model; Logistic = Logistic Regression Model.

LPM coefficients represent percentage point changes in delegation probability.

Logistic regression present odds ratios (OR); they confirm consistent effects across specifications

Showing that the effects remain robust when we reduce the sample to the subjects our theory predicts have reason and possibility to delegate in order to offload responsibility, supports our hypothesis that participants delegate in order to offload responsibility, rather than to merely create social distance. Of those subjects we theorize have reason to delegate, 50% in fact delegate in the Social-Image Game and 56.8% in the Self-Image Game.

2. Transfer Decisions

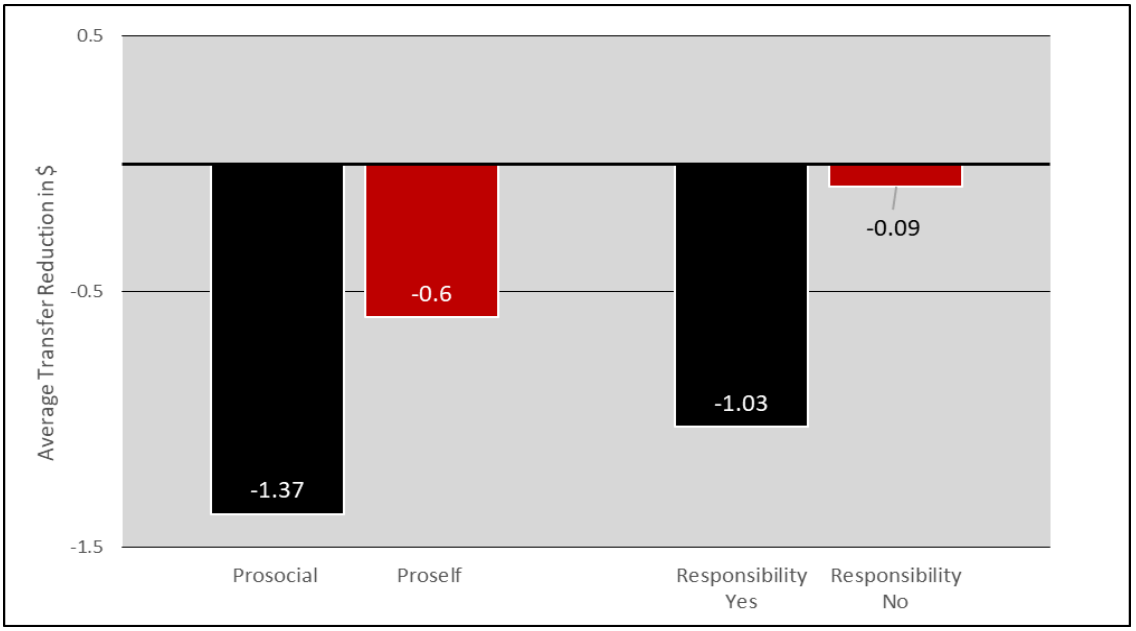
a) Prosociality

As prosocials should transfer more than proself Allocators without delegation, we hypothesize that the more prosocial an Allocator is, the greater the margin by which that Allocator will reduce their transfers compared to their transfer without delegation option. We posit:

Hypothesis D_4: ***The more prosocial an Allocator, the more should they reduce transfers relative to their transfer choices without option to delegate.***

First, we analyze the transfers of **prosocial** Allocators. We compare the transfers they made when they had no delegation option (recall that these transfers were incentivized and implemented with a probability of 1/10) with the transfers they made in the main game with the option to delegate to the AI. We find that prosocial subjects transferred $4.12 on average without a delegation option, and significantly less—$2.75—with the delegation option ($t(98) = -3.451$, $p = 0.001$). This gives prosocials a margin of $1.37 by which they reduce their transfers when delegating.

Figure 5. *Reduction in Transfer comparing Transfer with delegation option and without by Prosociality and Responsibility Attribution*



Second, we compare the transfers of **proself** Allocators without a delegation option ($2.83) to their transfers when they have a delegation option ($2.25). Proself types contribute significantly less without the delegation option ($t(98) = -3.451$, $p = 0.001$), but they reduce their

transfers significantly less than prosocial types, consistent with our hypothesis D_4. We illustrate the results in **Figure 5** below.

For the regression analysis we performed a mixed-effects Tobit model, as we compare transfer decisions within and across subjects. Transfers are censored at \$0 and \$10. The model includes the dependent variable "Transfer" and the independent variables "Responsibility" and "SVO" plus a dummy that distinguishes between the transfers in the *AI Treatment* with delegation option and the transfers without delegation option. We refer to the dummy as "AI Treatment". Finally, we include an interaction term of "Responsibility" and that dummy.

We run two models. In the first, we restrict the sample to subjects who delegated to the AI. Even though the sample is small, we expect the impact of prosociality to be the strongest, as delegating subjects are more prosocial than the rest of the sample (39.18 points versus 15.57). The second model analyzes the full sample.

As presented in **Table 8** below, results show first a general effect of prosociality: as Allocators become more prosocial, they tend to transfer more to the Recipient (β= 0.025, p< 0.01). Second, as hypothesized in D_4, the interaction term between "Prosociality" and "AI Treatment*"* is significant and negative (β= –0.011, p= 0.022), indicating that the more prosocial Allocators are, the more they reduce their actual transfers compared to their transfers without delegation option.

We conclude that Allocators strategically lower their transfer as delegation to AI and the individual level of their prosociality permit. Next, we test whether attributing more responsibility to the AI allows participants to reduce their transfers to a greater extent.

b) Responsibility Attribution

If Allocators exploit their delegation to the AI strategically, we expect the impact of the delegation on transfers to increase with the degree of responsibility that Allocators think they can offload to the AI.

The difference between Allocators' transfers when they have a delegation option and their transfers when they do not have this option should increase with the degree of responsibility they attribute to the AI. We posit:

Hypothesis D_5: ***The more* responsibility *subjects attribute to the AI, the more they reduce their transfers when they can delegate compared to their transfers without a delegation option.***

To report descriptive data, we first recode responsibility as a binary variable, distinguishing between Allocators who attributed responsibility to the AI and those who did not. We begin by analyzing the behavior of subjects who attributed responsibility to the AI in the Social-Image Game, comparing their transfers with a delegation option available (\$2.30) to their transfers without this option (\$3.33; t(54)=-2.3945; p= 0.02), yielding a difference of \$1.03.

Second, we analyze the transfers of subjects who did not attribute responsibility to the AI. Again, we compare their transfers with the delegation option available (\$3.26) to their transfers without this option (\$3.17; t(44)=0.199, p=0.84), yielding a difference of merely \$0.09. As hypothesized in D_5, subjects who attributed responsibility to the AI reduced their transfers

more strongly ($1.03) than participants who did not attribute responsibility to the AI ($-0.09). We illustrate the results in **Figure 5** above.

To statistically confirm this descriptive difference-in-differences analysis, we conduct Tobit regression models. We calculate results for two samples: first, restricted to Allocators who delegated to the AI, and second, the full sample of the AI Treatment. Results do not differ across the two models.

As hypothesized, the more responsibility Allocators assign to the AI, the more they reduce their transfers with delegation option compared to their transfers without delegation option (β= 0.019, p= 0.01).

The interaction effect between "Responsibility Attribution" and "Delegation Option" shows that responsibility attribution has a significantly stronger impact on transfers when subjects delegated, whereas in the condition without the delegation option—responsibility attribution has little impact on transfer amounts (β= 0.029, p= 0.02).

In the second model, we repeat the analysis with the full sample of the *AI Treatment* and find that the results are supported. The results for both models are shown below in **Table 8**.

Table 8: Lowering Transfers as a function of Responsibility & Prosociality

Variables Transfer	**(1) AI Treatment — All participants** Mixed-Effects Tobit (β, SE)	**(2) AI Treatment — Delegators only** Mixed-Effects Tobit (β, SE)
Delegation Option (=1; no condition =0)	-0.021 (0.217)	-0.356 (0.865)
SVO (Angle)	0.016** (0.004)	0.029* (0.010)
SVO × Delegation Option	-0.012** (0.004)	-0.027** (0.009)
Responsibility (scale 0-100)	0.019** (0.008)	-0.041*** (0.014)
Responsibility × Delegation Option	-0.029** (0.009)	-0.023*** (0.009)
Constant	3.038*** (0.206)	2.037** (0.882)
Censored left	43	15
Censored right	4	1
Observations	200	64
Subject clusters	100	32
Wald Chi2(5)	44.74	65.24***

Notes:
Significance: * p <0.10, ** p <0.05, *** p <0.01.
Mixed-effects Tobit censored at $0 and $10; robust SEs clustered by subjects.
Actual transfers in the AI Treatment (=1) are compared to counterfactual choices = 0.

Thus, the data suggests that, consistent with our hypothesis D_5, Allocators strategically lower their transfers to the extent that they can offload responsibility to the AI.

E. Norm-Beliefs

Finally, delegation should affect the attribution of responsibility and the amounts Allocators transfer, but should not affect participants' judgments of the transfer's fairness. The data support this hypothesis: Allocators' norm-beliefs were tightly clustered around an equal split. In the Self-Image Game, delegating Allocators reported that they ethically should transfer on average \$4.29 (n= 62), while non-delegating Allocators (n= 32) reported a similar \$4.56 ($t(98)= 0.936$, $p= 0.35$). In the Social-Image Game, delegators reported an average ethical norm of \$4.47 (n=32), compared to an almost equal \$4.43 (n= 68) among non-delegators ($t(98)= 0.063$, $p= 0.94$). Including norm-beliefs as covariates in our regressions that analyze delegation and transfer decisions yields no significant effects and does not alter the main results. Norm-beliefs also do not differ significantly by treatment (AI vs. LimeSurvey) in either the Self-Image or Social-Image Game.

F. Summary of Results

In sum, we find support for all four building blocks of our theory: First, more than a third of the participants changed their normative decision environment through delegation, such that they assume less responsibility for the transfer. Second, Allocators capitalize on the opportunity to delegate by reducing their transfers.

Third, Allocators are motivated by self-image concerns to delegate to and involve the AI even when they are not observed and their choices cannot be associated with them. Delegators also protect their social-image: almost 70% expect delegation to lead Observers to attribute less responsibility to them compared to when they decide themselves. These expectations drive their delegation decisions: those who expect Observers to attribute less responsibility to them are significantly more likely to delegate to the AI.

We also find that their expectations are correct: contingent on delegation, more than 70% of Observers perceive delegating Allocators as less accountable, showing that Allocators can often positively impact their social-image by delegation.

Fourth, and finally, Allocators act strategically and are behaviorally rational when considering delegation. Prosocials—who face higher moral costs from violating the norm and, absent a delegation option, would transfer more—are more likely to delegate than proselfs. And when they delegate, prosocials reduce their transfers more. Moreover, Allocators who expect delegation to lower the responsibility attributed to them are also more likely to delegate and, upon delegating, reduce transfers more than those who do not have that expectation.

V. Discussion

A. Internal Validity

1. Responsibility and Cognitive Dissonance

We asked Allocators about their responsibility attribution after they made their experimental choices to avoid the risk that the elicitation could influence Allocators' delegation and transfer choices. However, this raises the possibility that Allocators might have aligned their stated responsibility with the transfer choices they made to avoid cognitive dissonance

(Festinger, 1957; Cooper, 2012). To assess whether our responsibility results were biased, we compare Allocators' responsibility attribution in the AI Treatment to the responsibility attribution reported of the Observers in the same treatment. We then make the same comparison between Allocators and Observers in the LimeSurvey Treatment, which serves as a robustness check. There is no significant difference between responsibility attribution in the two roles in either treatment. This suggests that the perceived responsibility that Allocators indicated was not relevantly biased by cognitive dissonance, but rather reflects their largely unbiased belief about responsibility that may lead them to decide to delegate.

Our conclusion that Allocators' estimates were not distorted by their choices is further supported by the responsibility attribution of the Observers, who cannot benefit from that attribution, and who attributed even more responsibility to the AI than Allocators did.

The conclusion that Allocators reported presumptively unbiased responsibility values also is aligned with our theory of why Allocators delegate: Allocators will choose to delegate if they assume that delegation will reduce social-image and self-image costs, but these costs are only lowered if Allocators expect that Observers believe delegation will reduce the responsibility the Allocators assume, and that Allocators genuinely believe the same. Thus, Allocators have an incentive to accurately estimate the responsibility Observers attribute to them and the AI.

2. Alternative Motivations for Allocators' Delegation Choice

We hypothesize that Allocators delegate to reduce the responsibility they assume for the transfer, whether self-perceived or attributed to them by others. We rule out potential alternative motivations in two ways. First, we made specific design choices that should have rendered alternative motivations for delegating—like experimenter demand effect, desirability effects, or curiosity—unlikely. Second, we included the LimeSurvey Treatment, which, when compared to the AI Treatment, allows us to conclude that the observed treatment differences are driven by participants' intent to reduce their perceived responsibility for an unequal allocation.

Experimenter demand effects. If Allocators chose to delegate in an attempt to support what they believed to be the experimenters' research goals, our results would be contaminated by demand effects. To counter this possibility, we ensured that our research objective was not revealed to participants. We also delayed the debriefing process until the study was completed, in order to avoid potential communication between past and future subjects of different treatments. Subjects knew only their own treatment and could not be sure whether a particular action would help or hinder the researchers' goals compared to the unknown control group. We also employed a double-blind experimental design to increase the social distance between participants and experimenters. Participants were unaware of the experimenters' identities, reducing the likelihood that they would adjust their behaviour to help achieve whatever they presumed to be the experimenters' goals (Hoffmann et al., 1994).

We also test the delegation rate against our theoretical predictions. Specifically, we compare the delegation rate of subjects who both have other-regarding preferences and attribute responsibility to the AI with the rate of delegation among those who either lack other-regarding preferences or do not attribute responsibility to the AI. This allows us to distinguish delegation

consistent with our theory from delegation that is not—and that may instead be driven by other motivations, such as demand effects. By distinguishing, we ensure that if demand effects were present, the noise they introduce does not undermine the robustness of our results. This observation also applies to the alternative motivations discussed in the following paragraphs (we do not repeat this point there even though it applies)—such as intrinsic curiosity about the AI or social desirability effects. It also holds if subjects were to mistakenly assume that the AI provided substantive input to the transfer decision, despite the strict experimental protocol ruling out that the AI would signal any particular choice or transfer decision was better or morally preferable than another.

We find that delegation among those who meet the theoretical criteria is significantly more likely (see Section A). Moreover, our regression analysis shows that the motivation we theorize—offloading responsibility among those who would otherwise feel bound by fairness norms and thus contribute more—is driving both the decision to delegate and the level of transfers (see Sections C and D).

Desirability effects. We took care to make no statement or design choice that could communicate the message that delegation is morally or ethically desirable. To the contrary, delegation, for the strategic purpose of allocating less, is likely to be seen as a departure from a socially desirable course.

Motivation to engage with AI. We also sought to ameliorate the possibility that Allocators might decide to delegate out of curiosity, instead for strategic BSM reasons. The instructions made very salient to the subjects that the AI will follow a scripted transfer protocol and that the communication with the AI was scripted, such that subjects should be aware of what to expect of the AI. In addition, subjects completed both games with and without an Observer consecutively, suggesting that at least in the second game curiosity should not have relevantly motivated their delegation decision. And indeed, we do not find a substantial difference in delegation choices comparing the first and second game. Also, a curiosity motivation could not explain why Allocators were more likely to delegate the more responsibility they attributed to the AI and the more prosocial they were and why they lowered their transfers more as they attributed more responsibility to the AI.

Substantive Influence. Our aim was to demonstrate that subjects delegate in order to share responsibility, rather than because they want the AI to advise or persuade them that a self-interested transfer is acceptable. Therefore, we ensured that the subjects understood that the AI would strictly adhere to the scripted transfer protocol. This ensured that subjects were aware that the AI would initially suggest a transfer of $0.00 as a matter of protocol rather than as a matter of a moral recommendation. We tested participants' understanding of this point in the control questions. Even if the neutral protocol did nevertheless influence some participants, or if they did not want to progress through the protocol to make a higher transfer in order to save time or effort, this effect should also be present in the LimeSurvey control treatment and thus cancel out. This is particularly true since LimeSurvey carried out all participants' transfers in accordance with this protocol, not just those of participants who delegated as was the case in the AI Treatment.

Social Distance. Consistent with our BSM theory, we find that Allocators in our AI Treatment in both the Social-Image Game and the Self-Image Game transfer significantly less than in both games of the LimeSurvey Treatment.

In both treatments either the AI or LimeSurvey completes the final act in the causal chain by writing the message and providing the collection code to the Recipient. As a consequence, our design should hold social distance constant across the two treatments and thus cancel it out as a cause of the lower transfers in the AI Treatment.

However, one might argue that, in participants' perception, the AI may increase perceived social distance to a larger degree than LimeSurvey does because of its greater complexity and quasi-agency, even though the actual actions and interference with Allocators' autonomy are identical for both AI and LimeSurvey. Participants may perceive the AI as merely generating social distance, in a similar way to a well-known trolley dilemma. In this scenario, individuals are more likely to push a man onto the track to prevent the train from killing a larger group of people if they can pull a lever to push him onto the track rather than push the man by touching him directly. If participants were to delegate to the AI to increase social distance, perceiving the AI as a special social distance-conveying instrument to ease their responsibility and make it morally less costly for them to make a low transfer, this would still be a BSM effect, which we will lay out in the next section.

But notice, perceiving the AI as a mere instrument to create social distance would suggest a reduction in the responsibility that participants attribute to themselves, it would not imply that the AI itself carries any responsibility for the transfer (any more than the lever does in the trolley problem). Thus, when both Allocators and Observers attribute responsibility for the transfer to the AI and our data show that this attribution appears to motivate Allocators' delegation decisions and leads them to transfer less, our data appears to suggest that Allocators view the AI as an entity that can assume moral responsibility rather than a mere (complex) process interfering with their agency. And even if their perception is incorrect, the fact that this perception is widespread among both Allocators and impartial Observers shows its relevance to people's moral decision making.

B. External Validity

1. Lab Population and Field Evidence

Our participant sample consists mostly of NYU students and externals who were recruited through NYU Stern's School of Business Behavioral Lab and the lab of the NYU Economics Department. While externals and students do not appear to exhibit different behaviors in our study, our subjects received more academic training compared to a representative sample, which might increase the sophistication of our subjects in self-managing their moral behavior.

Also, our subjects are comparatively young, and the literature suggests that individuals exhibit more prosocial behavior with age. But this young subject pool may work against the BSM effect we find. Our SVO results show that with increasing prosociality participants are more likely to delegate to the AI, and if prosociality increases with age, subjects may also become more prone to using BSM with age. Also, people may become more effective with age at managing their moral costs and thus be more likely to use BSM as one of the tools they apply

to moral decision making. On the other hand, they might not be as accustomed to using AI tools.

Evidence that indirectly supports our claim that people should be sophisticated enough to 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 in the lab like ours (e.g., Freddi 2021).

2. The External Validity of Economic Games

One might question our results by noting the relatively low stakes in our dictator games and argue that, at higher stakes, individuals would simply pursue self-interest and endure the moral costs for self- and social image. While people still might delegate to the AI to reduce their felt responsibility, it would not change their transfer choices. Yet, evidence from poorer countries—where identical nominal amounts represent far larger real gains show that patterns of social preferences remain strikingly stable: Cameron (1999) reports ultimatum game results from Indonesia with stakes equivalent to wages of multiple months; Henrich et al. (2005 & 2010) set stakes in dictator games at one-day wages and Andersen et al. (2011) offer stakes of up to the equivalent of 1,600 working hours. Carpenter et al. (2005) and List and Cherry (2008) provide stakes of $100. Munier and Zaharia (2002) increased stakes up to 2,000 French Francs. While higher stakes reduced transfers, prosocial behavior in dictator and ultimatum games turned out to be robust across a wide range of stake sizes, so the motive to engage in BSM should persist—and might even intensify—when stakes are higher. Thus, even when stakes are higher, the delegation might make people's allocation choice more self-interested than it would be otherwise.

Rather than being an obstacle, the modest payments may make the experimental test more stringent: we find that participants are willing to delegate their transfer decision even for small rewards, and larger incentives may increase that willingness.

C. Implications for AI Regulation: Human-in-the-Loop, Wiggle Room and the Responsibility Gap

Our findings have a number of implications for the regulation of AI. The first relates to our understanding of the goals that AI regulation should be pursuing.

The EU AI Act, officially the Regulation of the European Parliament and of the Council on Artificial Intelligence, is a comprehensive framework for regulating AI systems within the European Union. It was approved by the EU Council on May 21, 2024, and entered into force on August 1, 2024. The Act aims to ensure AI systems are safe, transparent, and respect fundamental rights while fostering innovation and economic growth. In the future, the EU AI Act may well serve as a model for AI regulations outside the EU, including in the U.S.

A cornerstone of the EU AI Act is the so-called "human-in-the-loop" (HITL) requirement. This is laid out in Article 14 of the Act, which obliges providers of high-risk AI systems to ensure that natural persons can "understand, detect anomalies, avoid over-reliance, interpret, and override or stop the system." (Art. 14b). The logic is straightforward: human oversight serves as a safeguard against harm from erroneous or biased algorithmic outputs.

Complementary provisions reinforce this framework. Providers must supply transparency information to users about limitations and performance conditions (Art. 13), while deployers must assign competent personnel to exercise oversight, monitor the operation of the system, and—where they control inputs—ensure those inputs are appropriate and representative (Art. 26). Our results suggest that the HITL principle, while addressing some problems, may fail to reach or even worsen others. Specifically, while the HITL structure is supposed to address the inherent dangers of AI systems—such as opacity, bias, or technical failure—it does not address the risk we have analyzed in our study: that humans may exploit AI strategically for self-interested ends by offloading their responsibility to it. And there is reason to believe that the risk of responsibility offloading may loom large in the set of risks created by AI.

For example, in medical contexts, this moral offloading risk may be particularly acute. Consider a doctor using an AI system to interpret magnetic-resonance images (Bigman & Gray, 2018). The doctor does not simply accept or reject a final recommendation; she sets key assumptions during the process—such as which surgical options are feasible with the hospital's equipment and how to weigh patient-specific risks—and the system then returns recommendations with estimated probabilities of outcomes. If the doctor has financial or reputational incentives to favor surgery, the probabilistic output creates a convenient rationale. By emphasizing the model's quantified risks of non-surgical options, she can present the choice as "data-driven" while offloading responsibility for adverse outcomes to the system—a deflection that becomes easier the stronger the perception of the AI system's moral agency. This may be particularly likely when there is wiggle room; i.e. when there are different treatment options and no single option is judged by an external authority, such as the applicable medical practice guidelines, as negligent and invoking liability if outcome risks are realized. The doctors may offload responsibility to the AI for choosing within that wiggle room the treatment most lucrative for themselves, much as people exploit "moral wiggle room" in other domains by choosing ignorance or ambiguous justifications to serve their own interests (Grossman, 2014; Vu et al., 2023; Dong & Bocian, 2024; tho Pesch & Dana, 2024; Offer et al., 2024).

The AI Act addresses this form of misuse only indirectly. Article 14 requires that systems be designed to avoid over-reliance, but it does not explicitly address opportunistic incentives to "strategically over-rely" and exploit the AI as a moral alibi both by users following their self-interest and deployers like companies who want to benefit from users' self-interested actions. Article 26 requires deployers to ensure input data is appropriate, which constrains blatant manipulation of inputs, so it regulates the doctors' inputs in our example, but it does not directly regulate strategic framing of the AI's perceived moral agency. Liability provisions, including the proposed EU AI Liability Directive, ensure that harm can be traced back either to the provider (for defective systems) or the deployer (for improper use). Yet, these instruments focus only on closing the legal responsibility gap—that is, who is responsible for the design of the technology and oversight of its use (Matthias, 2004; Santoni de Sio & Mecacci, 2021). However, even if salient to users and deployers, the regulation may fail to guide behavior, if the psychological responsibility gap that is created when humans deliberately offload responsibility to AI to pursue selfish objectives undermines compliance, in particular where monitoring is unlikely or performed by a party interested in enticing the misconduct. Paradoxically, then, the very presence of a human in the loop can create new opportunities for moral evasion: instead

of acting as an error-correcting safeguard, the overseer may use the system's outputs to avoid responsibility for their own agency.

There are potential policy interventions that could address, at least in part, the psychological responsibility gap we demonstrate here. The most direct intervention would be to remove AI's capacity to perform the final causal act in settings in which the user may wish to diffuse responsibility. In our medical example, for instance, both doctors and AI may file independent recommendations, a process that has been found to be more effective compared to other forms of human-AI interaction. Such a policy would, in effect, constitute an expanded and re-focused version of the human-in-the-loop principle—the policy would mandate a human in the loop to perform the last causal act, in order to reduce responsibility offloading.

However, designing such a direct legal intervention would be exceedingly tricky. There is surely a countless variety of potential transactions in which responsibility offloading through AI intervention may occur, and each may constitute specific risks and benefits of human-AI interaction. In our medical example, for instance, doctors' judgments may at some point become more a source of error than improvement. Therefore, legal intervention would either have to develop gradually by specifying contexts in which the expanded HITL rule would apply, or, would have to take the form of an all-encompassing standard. Slowly building out a set of rules would likely require Congress to empower an expert agency—perhaps the Federal Trade Commission, if it survives the current attack on its constitutional status. But imposing a standard—for example, imposing negligence liability on an AI developer that fails to reasonably anticipate responsibility offloading in a particular setting or to guard against it by requiring the human to perform the last causal act—is likely to impose significant costs in balancing possible negative effects of anticipated responsibility offloading with the benefits of a more agentive role of AI in particular cases.

A related policy intervention may require an AI to disclose information designed to blunt the responsibility offloading effect—for example, information making clear that the AI has no autonomous will and that it is performing the final causal act under orders from the principal. Our experiments do not test the effect of disclaiming language, but given our results such testing would be beneficial to assess the viability of this type of intervention. Note, however, that disclaimer interventions face the same difficulty that bans on AI "final causal act" intervention do—policymakers would need to understand costs and benefits of AI in a whole range of human-AI interactions in which responsibility offloading may be facilitated by AI in order to mandate disclosure effectively.

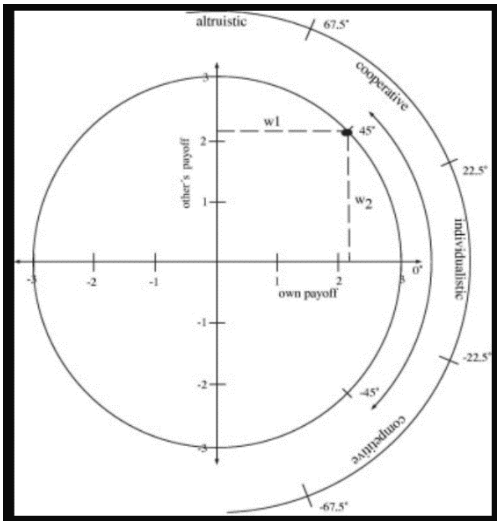
Appendix

I. Methods - *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] 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]

Figure 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 use the binary distinction between proself and prosocial types, it should be understood that most individuals classified as proself will also have other-regarding preferences of some degree, while individuals classified as prosocial 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 delegate in order to reduce their

transfers and moral costs. However, we expect the likelihood to delegate 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 when pursuing 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).

References

Andersen, S., Ertac, S., Gneezy, U., Hoffman, M., & List, J. A. (2011). Stakes matter in ultimatum games. *American Economic Review,* 101(7), 3427–3439.

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.

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., & 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.

Argenton, C., Potters, J., & Yang, Y. (2023). Receiving credit: On delegation and responsibility. *European Economic Review, 158*, Article 104522.

Balliet, D., Parks, C., & 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.

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes*, 94(2), 74–85.

Bartling, B., & **Fischbacher, U.** (2012). Shifting the blame: On delegation and responsibility. *The Review of Economic Studies,* 79(2), 67–87.

Bernstein, M. H., Sheppard, B., Bruno, M. A., Lay, P. S., & Baird, G. L. (2025). *Randomized Study of the Impact of AI on Perceived Legal Liability for Radiologists*. NEJM AI, 2(6).

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition,* 181, 21–36.

Bonnefon, J.-F., **Rahwan, I., & Shariff, A.** (2024). The moral psychology of artificial intelligence. *Annual Review of Psychology,* 75, 1–26.

Cameron, L. A. (1999). Raising the stakes in the ultimatum game: Experimental evidence from Indonesia. *Economic Inquiry,* 37(1), 47–59.

Candrian, C., **& Scherer, A.** (2022). Rise of the machines: Delegating decisions to autonomous AI. *Computers in Human Behavior,* 134, 107308.

Carpenter, J. P., Verhoogen, E., & Burks, S. V. (2005). The effect of stakes in distribution experiments. *Economics Letters,* 86(3), 393–398.

Charness, G., & Gneezy, U. (2008). What's in a name? Anonymity and social distance in dictator and ultimatum games. *Journal of Economic Behavior & Organization*, 68(1), 29-35.

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

Custers, B., Lahmann, H., & Scott, B. I. (2025). From liability gaps to liability overlaps: Shared responsibilities and fiduciary duties in AI and other complex technologies. *AI & Society,* 40(5), 4035–4050.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114–126.

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.

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.

Dong, M., & Bocian, K. (2024). Responsibility gaps and self-interest bias: People attribute moral responsibility to AI for their own but not others' transgressions. *Journal of Experimental Social Psychology,* 111, 104584.

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.

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.

Feier, T., Gogoll, J., & Uhl, M. (2022). Hiding Behind Machines: Artificial Agents May Help to Evade Punishment. *Science and Engineering Ethics,* 28, 19.

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.

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.

Gächter, S., & Renner, E. (2010). The effects of (incentivized) belief elicitation in public goods experiments. Experimental Economics, 13(3), 364–377.

Gawn, G., & Innes, R. (2019). Who delegates? Evidence from dictator games. *Economics Letters,* 181, 186–189.

Gawn, G., & Innes, R. (2021). Machiavelli preferences without blame: Delegating selfish vs. generous decisions in dictator games. *Journal of Behavioral and Experimental Economics,* 90, 101615.

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

Gill, T. (2020). Blame it on the self-driving car: How autonomous vehicles can alter consumer morality. *Journal of Consumer Research,* 47(2), 272–288.

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

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds & Machines,* 30, 99–120.

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.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., et al. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences,* 28(6), 795–815.

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., et al. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *Science,* 327(5972), 1480–1484

Hill, A. (2015). Does delegation undermine accountability? Experimental evidence on the relationship between blame shifting and control. *Journal of Empirical Legal Studies,* 12(2), 311–339.

Hoffman, E., McCabe, K., Shachat, K., & Smith, V. (1994). Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior,* 7(3), 346–380.

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

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

Kaminski, M. E & Selbst, A. D, (2025). An American's Guide to the EU AI Act. *University of Colorado Law Legal Studies Research Paper* No 18-25.

Köbis, N. C., Bonnefon, J.-F., & Rahwan, I. (2021). Bad machines corrupt good morals. *Nature Human Behaviour, 5*, 679–685.

Kneer, M., & Christen, M. (2024). Responsibility gaps and retributive dispositions: Evidence from the US, Japan and Germany. *Science and Engineering Ethics,* 30(6), 51.

Leib, M., Köbis, N., & Soraperra, I. (2025). Does AI and human advice mitigate punishment for selfish behavior? An experiment on AI ethics from a psychological perspective *(arXiv:2507.19487). arXiv. https://doi.org/10.48550/arXiv.2507.19487*

Lima, G., Grgić-Hlača, N., & Cha, M. (2021). Human perceptions on moral responsibility of AI: A case study in AI-assisted bail decision-making. In *Proceedings of the Conference on Human Factors in Computing Systems* (1–17).

List, J. A., & Cherry, T. L. (2008). Examining the role of fairness in high stakes allocation decisions. *Journal of Economic Behavior & Organization,* 65(1), 1–8.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes,* 151, 90–103.

Matthey A. & Regner T. (2015), More Than Outcomes: The Role of Self-Image in Other-Regarding Behavior. *Review of Behavioral Economics*, 2(4), 353-378.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology,* 6(3), 175–183.

Monaco, L. (2024, March 7). Remarks on artificial intelligence enforcement and corporate crime. *U.S. Department of Justice press release.*

Munier, B., & Zaharia, C. (2002). High stakes and acceptance behavior in ultimatum bargaining. *Theory and Decision,* 53(3), 187–207.

Munn, L. (2023). The uselessness of AI ethics. *AI and Ethics* 3(3), 869–877.

Nyholm S. (2018). Attributing Agency to Automated Systems: Reflections on Human-Robot Collaborations and Responsibility-Loci. *Science and Engineering Ethics*, 24(4), 1201-1219.

Oexl, R., & Grossman, Z. J. (2013). Shifting the blame to a powerless intermediary. *Experimental Economics,* 16(3), 306–312.

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

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.

Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology,* 34(4), 1057–1084.

Simmler, M. (2024). Responsibility gap or responsibility shift? The attribution of criminal responsibility in human–machine interaction. *Information, Communication & Society,* 27(6), 1142–1162.

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

Slonim, R., & Roth, A. E. (1998). Learning in high stakes ultimatum games: An experiment in the Slovak Republic. *Econometrica,* 66(3), 569–596.

Steffel, M., Williams, E. F., & Perrmann-Graham, J. (2016). Passing the buck: Delegating choices to others to avoid responsibility and blame. *Organizational Behavior and Human Decision Processes,* 135, 32–44.

Stuart, M. T., & Kneer, M. (2021). Guilty artificial minds: Folk attributions of mens rea and culpability to artificially intelligent agents. *arXiv (arXiv:2102.04209).* https://doi.org/10.48550/arXiv.2102.04209.

Tontrup, S., Arlen, J. & Sprigman, C. J. (2025a). Behavioral Self-Management and the Strategic Shifting of Fairness Norms. *NYU Working Paper.*

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.

Thielmann I., Spadaro G., & Balliet D. (2020). Personality and prosocial behavior: a theoretical framework and meta-analysis. *Psychological Bulletin,* 146, 30–90.

tho Pesch, F., & Dana, J. (2024). Attributional ambiguity reduces charitable giving by relaxing social norms. *Journal of Experimental Social Psychology,* 110, 104530.

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., Joireman, J., Parks, C. D., Van Dijk E. (2013). The psychology of social dilemmas: a review. *Organizational Behavior and Human Decision Making Processes*, 120, 125–41.

Vu, L., Soraperra, I., Leib, M., van der Weele, J., & Shalvi, S. (2023). Ignorance by choice: A meta-analytic review of the underlying motives of willful ignorance and its consequences. *Psychological Bulletin,* 149(9–10), 611–635.