THE EFFECTS OF INVESTIGATIVE SANCTIONING SYSTEMS ON WRONGDOING, REPORTING, AND HELPING: A MULTI-PARTY PERSPECTIVE

ONLINE APPENDIX

The following document is the appendix for the paper "The Effects of Investigative Sanctioning Systems on Wrongdoing, Reporting, and Helping: A Multi-Party Perspective". The appendix contains descriptions of the methods and findings of two supplementary studies; details on the methodology of our main studies 1 and 2, and some additional analyses and results for our main studies 1 and 2. The appendix contains the following below:

APPENDIX A: COMPANY CODES OF CONDUCT TEXTUAL ANALYSIS STUDY

Study Description and Sample

To explore the empirical prevalence of investigative sanctioning system and the extent to which companies make such systems salient to potential and current employees, we conducted an exploratory text analysis in which we had two research assistants code the codes of conduct of the 100 largest U.S. employers (according to Fortune 500, 2017).[1] In addition, the assistants also coded the types of behaviors that were considered to be misbehaviors in the codes of conduct. The goal of this part of the analysis was to assess to what extent problematic individual-level interpersonal behaviors (e.g., mistreatment) are the target of companies' investigative sanctioning system.

Coder Instructions and Description of Coding Categories

The research assistants first downloaded the codes of conduct (or codes of ethics for some companies) from the (parent) company's website. They then manually coded each of the downloaded documents according to the categories and definitions described below. Specifically, the two coders received the following instructions and coded the company documents according to the following category definitions:

General Instructions to Coders

We are interested in coding for the presence of formal surveillance and sanctioning systems in organizations, which broadly can be defined as "official systems and/or processes that can be documented and verified by an independent observer". Thus, you as coders take the role of the independent observer and try to verify in the companies' codes of conducts whether such systems exist. Below we provide definitions for the different categories you should code for.

Mentions formal reporting systems

- Code "Yes" (=1) for formal reporting systems if there is a description in the code of conduct about any sort of program, system, procedure, or process on how employees can report misconduct.
- Program or system examples would be hotlines, ethics lines, web portals, etc.
- Procedures or processes examples would include guidelines to report to a supervisor or to report to a specific manager or ethics officer or ombudsperson.

Mentions formal internal investigative procedures

- Code "Yes" for internal investigative procedures if there is a description in the code of conduct about any sort of procedure or process that the company will engage in to investigate and detect any allegations or any suspected wrongdoing.
- Examples would be descriptions about launching an internal investigation (or conducting an inquiry or audit) if misconduct is suspected or reported (e.g., by an employee or a client).

Mentions formal external investigative procedures

- Code "Yes" for external investigative procedures if there is a description in the code of conduct about engaging any sort of third-party investigations such that the organizations hires outside organizations or investigative committees that are external to the company to investigate suspected misconduct.
- Examples would be any description about the possibility of launching third-party investigations if misconduct is suspected or reported.
- Simply stating that the police (or some other law enforcement agency) will be informed about criminal behavior does not count. Similarly, a routine external / independent audit, does not count as an external investigative procedure.

Mentions formal sanctioning systems

- Code "Yes" for formal punishment or sanctioning systems if there is a description in the code of conduct about any sort of program, system, procedure, or process that the company will follow that directly associates ethical and unethical behavior with formal rewards and punishments.

---

[1] We excluded "Yum China Holdings" as most of their 450,000 employees are not based in the U.S. and "Publix Super Markets", as we could not find a publicly accessible code of conduct.

- Examples would be any descriptions about the company's disciplinary processes if an employee is found guilty of misconduct (e.g., descriptions about the company firing or having "zero-tolerance" or punishing employees if they are found to commit misconduct) or any descriptions about the company's possible rewards for behaving desirably (e.g., descriptions about the company offering a reward or bonus or part of performance evaluation for reporting misconduct).
- Other examples for the evidence of formal sanctioning systems would be descriptions about systems that offer possible rewards and punishments (e.g., evaluations, promotions, salary, and bonuses) related to ethically relevant employee behaviors.

Mentions interpersonal violations as unethical and inappropriate conduct

- Code "Yes" for interpersonal violations if the codes of conduct describe interpersonal behaviors (i.e., behavior that specifically harms another individual or other individuals in the organization) as inappropriate or unethical or harmful to the organization or that such behavior should be avoided and / or punished.
- Examples of interpersonal misbehaviors are unfair treatment, harassment, discrimination, abusive behavior, bullying, incivility, etc.

Mentions non-interpersonal violations as unethical and inappropriate conduct

- Codes "Yes" for non-interpersonal violations if the organization describes non-interpersonal behaviors (i.e., behavior that is more likely to harm the overall organization as opposed to specific individuals in the organization) as unethical or inappropriate or harmful to the organization or that such behavior should be avoided and / or punished.
- Examples include fraud or stealing from the organization, corruption, bribery, etc.

Results

Coding Reliabilities. To estimate coding reliability, 41 of the 100 documents were coded by both research assistants. Both coders found that reporting systems and inappropriate non-interpersonal behaviors (fraud, bribery, etc.) were mentioned in all codes of conduct and thus there were no disagreements between the two coders at all for these two categories. For the other coding categories, reliabilities, as measured by Krippendorff's alpha, were high (internal investigations: α=1.00; external investigations: α=.79, inappropriate interpersonal behavior: α=1.00) with the exception of formal sanctioning systems. For formal sanctioning systems, there was one of 41 cases ("Best Buy") in which the two coders did not agree. As both coders found that a formal sanctioning system was mentioned in all other 40 cases, this disagreement in one case is enough to lead to an unsatisfactory reliability according to Krippendorff's alpha (α=0.00). However, given the extreme distribution of coding values, we nevertheless report the results below.

Coding Frequencies. The frequency of codings for each category were the following: formal reporting systems: 100.0%, internal investigations: 99.0%, external investigations: 5.5%, formal sanctioning systems: 98.5%, interpersonal misbehaviors: 95.0%, non-interpersonal misbehaviors: 100.0%. When the two coders disagreed (which occurred only in two cases in total), we did not resolve this disagreement, but we simply report the average coding. The percentage values thus show how often a coder indicated that the respective category was mentioned in a code of conduct.

Discussion

The results of our exploratory content analysis of 100 codes of conduct of the largest U.S. employers show that almost all companies in our sample mention formal systems and procedures for investigating and eventually sanctioning employee wrongdoing in their codes of conduct. We can thus conclude that such systems are highly prevalent and that they are made salient to employees in the relevant company documents. Moreover, we also find that these systems not only target problematic behaviors at the organizational level (e.g., fraud, bribery, etc.) but that they are also used to detect and prevent problematic interpersonal behaviors at the employee level that are targeted at another individual and occur between employees within a company, suggesting that preventing individual-level harmful behaviors in the workplace is also highly important to organizations.

APPENDIX B: EMPIRICAL ASSESSMENT OF THE UNETHICALITY OF PARTICIPANT A'S BEHAVIORAL CHOICE IN STUDIES 1 AND 2

Study Description and Sample

Following Derfler-Rozin, Moore, & Staats (2016), we conducted a short validation survey using Prolific Academic (N=100, 74% female, M_{age} = 34.40, SD = 10.36). We used six scenarios from Derfler-Rozin, Moore, & Staats (2016) who had adapted a set of questions from a dissertation-based field study about ethics from Martin (2015), and we created six additional scenarios for a total of 12 scenarios to establish in a sample of adults that Participant's A's choice to increase his/her own points by taking away points from Participant B is perceived as 'inappropriate'.

Methods

Similar to Derfler-Rozin et al., (2016), we asked participants to rate several scenarios in terms of their appropriateness on a 9-point scale (1= completely inappropriate; 9 = completely appropriate). The six scenarios from Derfler-Rozin, Moore, and Staats (2016) were based on real-life contexts, in which three were filler scenarios that depicted ethically neutral behavior (e.g., delaying buying a birthday card for a friend by one day) and three represented inappropriate behavior (e.g., underreporting one's taxes).

The six scenarios that we created were based on abstract descriptions of monetary transactions with other parties in which three were filler scenarios (e.g., allocating 10 dollars each to two people) and three scenarios represented inappropriate behavior (e.g., being asked to split 20 dollars between oneself and another and choosing to transfer only 4 pounds to the other party while keeping 16 pounds for oneself). These three 'inappropriate' scenarios included our focal scenario: "*You have 10 dollars and the other person has 10 dollars (you both have 10 dollars each). You can decide whether to increase your own pool of money at the expense of the other person. If you decide to increase your own money, you will end up with 14 dollars and the other person will end up with 4 dollars only (your gain is not symmetrical to the other player's loss). Alternatively, if you decide to not increase your pool of money, you and the other person will each keep your 10 dollars each. You make a decision and you decide to increase your own earnings to 14 dollars and thus decrease the other person's earnings to 4 dollars. How appropriate is this decision*?".

Results

The mean rating of the "filler" real-life based scenarios (M=6.72, SD=0.87) and the "filler" games-based scenarios (M=6.89, SD=1.18) were not significantly different from each other, $t(99)=1.40$, $p = .16$). The mean rating of the "unethical" real-life based scenarios (M=2.98, SD=1.33) versus the "unethical" games-based scenarios (M=3.16, SD=1.45) were also not significantly different from each other ($t(99)=0.94$, $p=.35$).

Next, the mean rating on all six of the "filler" scenarios (M=6.80, SD=0.85) differed significantly from the mean rating of the inappropriate ones (M=3.07, SD=1.03, $t(99)=26.34$, $p<.001$). Lower ratings represent more inappropriate behavior. Importantly, the mean rating for choosing to asymmetrically increase one's own resources by decreasing another's resources in a monetary transaction (our focal behavior) (M=2.95, SD=1.93) was not significantly different from the mean rating for the three real-life "unethical" scenarios (M=2.98, SD=1.33, $t(99)=0.15$, $p=.88$), and it was significantly lower than the mean rating for the "filler" real-life based scenarios ($t(99)=16.95$, $p<.001$) and the mean rating for the "filler" games-based scenarios ($t(99)=16.77$, $p<.001$). For example, our focal behavior was perceived to be similarly inappropriate as underreporting one's taxes (M=2.61, SD=1.69, $t(99)=1.42$. $p=.16$). These findings suggest that our focal behavior of interest is perceived to be as unethical as committing tax fraud. The scenarios used are available upon request.

APPENDIX C: STUDIES 1 AND 2 EXPERIMENT MATERIALS

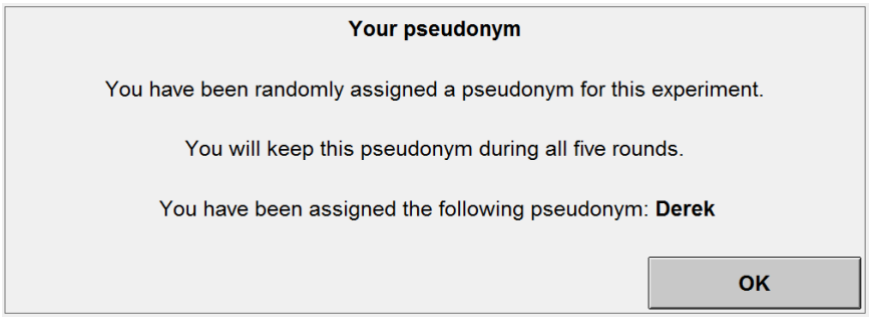
C1. Decision screens

Below we reproduce screenshots of the experimental software that show how participants interacted during the experiment and how they could enter their decisions into the computer. Participants received the same screenshots and read the same explanations before the experiment started.

HOW TO ENTER YOUR DECISIONS DURING THE EXPERIMENT

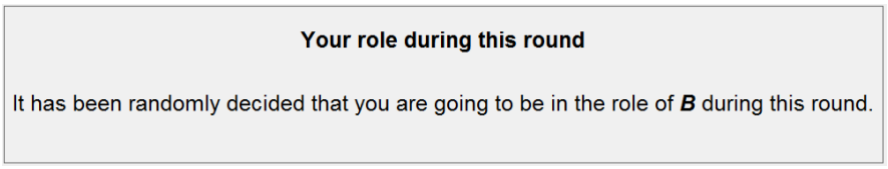
In the following, we show you in detail how you will make your decisions using the computer during the experiment.

At the very beginning of the experiment, a pseudonym is assigned to you:

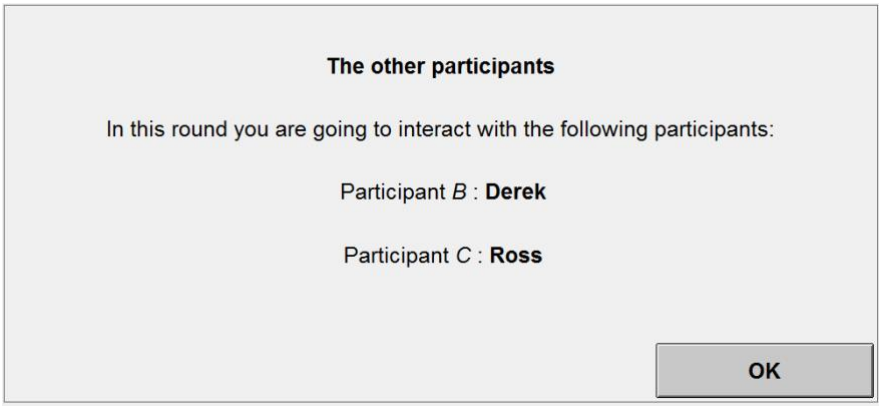


Click "OK" at the bottom right to move forward.

In each round, participants in the roles of *B* or *C* are informed of the role they will take in this round:

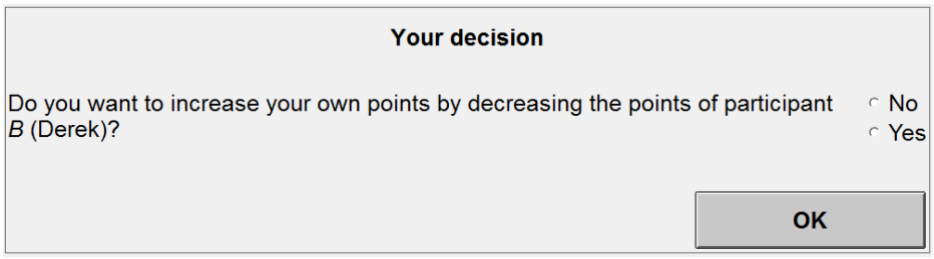


In addition, at the beginning of the round you will also learn the pseudonyms of the other participants with whom you will interact during this period:

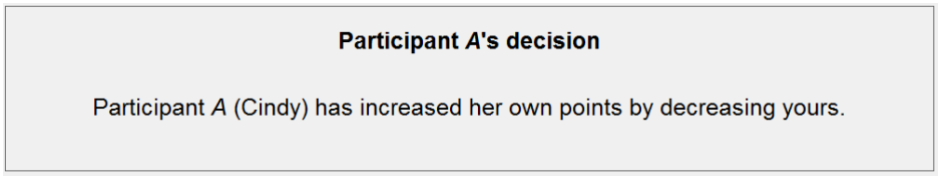


Above, you see the screen of participant *A*. The screens of *B* and *C* are equivalent.

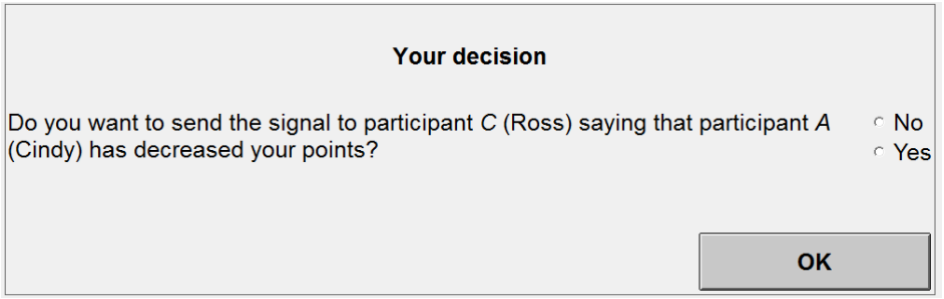
Participant *A* makes the first decision of each round. He / she can decide if he / she wants to increase his / her own points by decreasing *B*'s points. A makes this decision on the following screen:



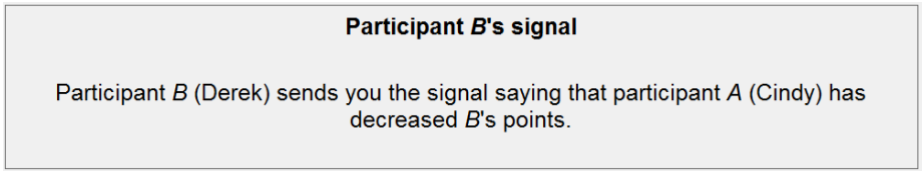
B is directly informed of *A*'s decision, i.e. *B* is informed whether his / her points have been decreased by *A*:



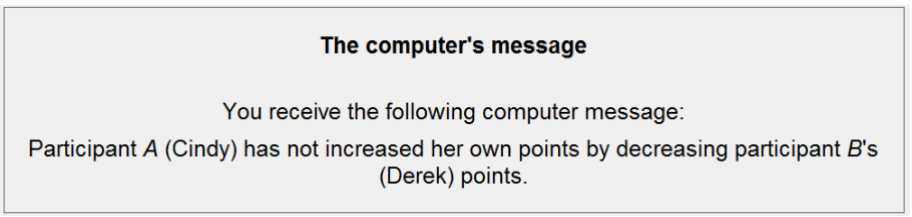
Then, *B* makes the decision if he / she wants to send the signal to *C* that *A* has decreased his / her points. You see *B*'s decision screen below:



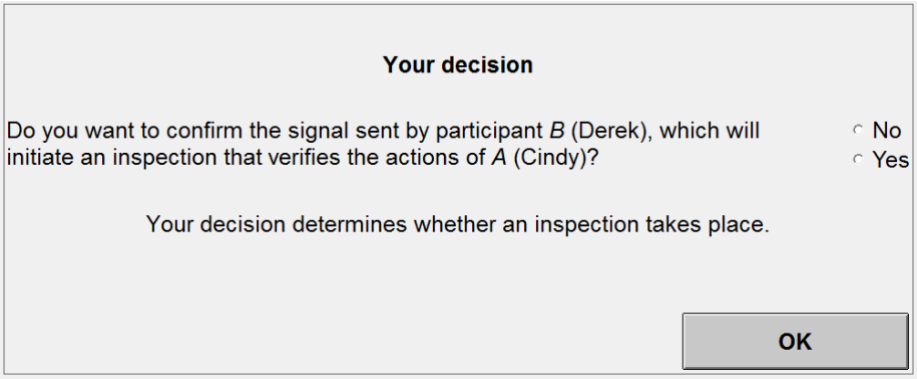
C then learns whether *B* has decided to send the signal to him / her. You see below the example where *B* decided to send the signal:



Before making the decision whether to confirm the signal, *C* also receives the computer's message (which is correct only 80% of the time) providing him / her with partial information about what *A* has done in the first step. The message will take, for example, the following form:



If *B* has sent the signal, *C* then makes the decision whether he / she wants to confirm *B*'s signal. *C* makes his / her decision using the following screen:



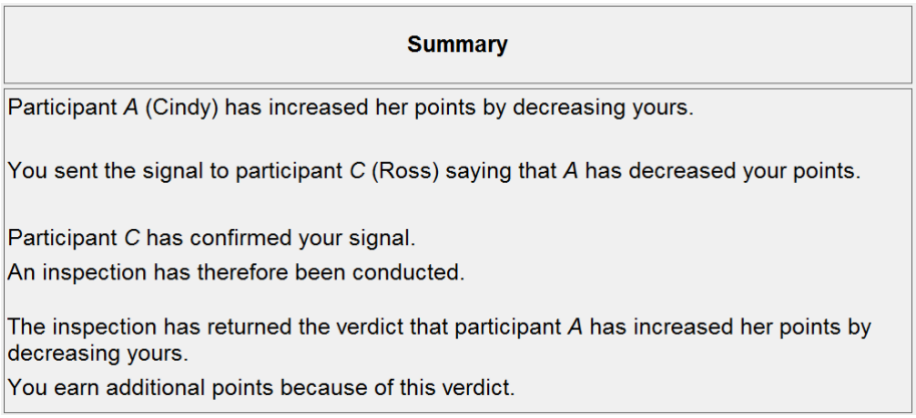
If *B* does not send a signal, *C* does not have a decision to make.

C's eventual decision is always the last of the round. If an inspection takes place, everyone learns the inspection's result on the last screen of each round.

This screen also provides a summary of what happened during this round and informs each participant of the profit he / she has made.

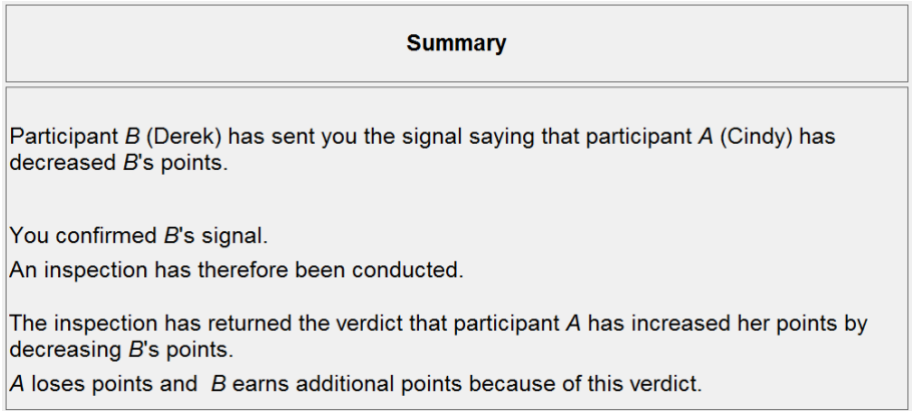
The content of the summary differs for the different participants. Everyone learns the result of the inspection, if an inspection has taken place, but while *A* and *B* know what A did in the first step (i.e., whether *A* has or has not increased his / her own points by decreasing *B*'s points), participant *C* will never know with certainty what truly happened and will never be informed of this decision, not even at the end of the round.

Here is an example of a summary for participant *B*:

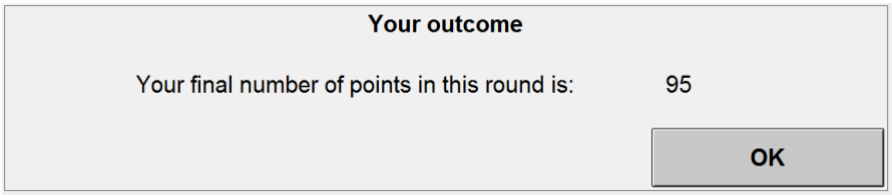


The summary that *A* receives is equivalent and contains the same information.

The summary for *C*, in contrast, is a little different. It does not contain information about the decision that *A* actually took in the first step. You can find an example of a summary for Participant *C* below:



At the bottom of the last screen, each participant is informed of the profit he / she has made in this round:



Once everyone has clicked "OK" on the summary screen, a new round begins in which you are randomly assigned to other participants.

C2. Summary of Experimental Instructions

The following summary of the instructions for the experimental game was read aloud to participants right before the game started:

All conditions:

Participant A makes the first decision. He/she can decide whether or not to increase his/her own points by decreasing B's points.

Independently of A's decision, B can decide to signal, or not, to C that A has decreased his/her points. If B sends this signal, C can decide to confirm it or not.

Before making this decision, C receives a message from the computer giving him/her an indication about what A has done in the first stage. This message is correct with a probability of 80% and it is false with a probability of 20%.

…

Surveillance and sanctioning system conditions only

…

If B sends the signal and C confirms it, an inspection takes place. The inspection renders a verdict about A's action in the first stage.

If the inspection finds that A has decreased B's points, A will lose points because of this verdict, and B will gain supplementary points. If the inspection finds that A has not decreased B's points, A will not lose any points, and B will not gain any supplementary points.

Remember that the inspection is capable of determining correctly A's actual action in the first stage in [100% vs. 85% vs. 55%] of all cases.

No system condition only

…

If B sends the signal, C's decision to confirm it or not leads to a verdict about A's action in the first stage.

If C confirms the signal, the verdict is that A has decreased B's points. A will lose points because of this verdict, and B will gain supplementary points If C does not confirm B's signal, the verdict is that A has not decreased B's points. A will not lose any points, and B will not gain any supplementary points.

APPENDIX D: EXPLORATORY ANALYSES (STUDIES 1 AND 2)

False Accusations in Study 1 and 2

Our experimental game contained the possibility of recipients falsely accusing actors of stealing when the actors had not actually done so. This is an important element of our game, as it is the possibility of such false accusations that creates uncertainty about the factualness of any given recipient report that is at the heart of the problem of detecting and punishing wrongdoing in situations where the wrongdoing cannot be perfectly observed. Below we analyze the prevalence of false accusations in our experimental conditions. A false accusation is defined as a recipient sending a report that the actor has stolen, when the actor has not actually done so.

Recipient False Accusations Study 1. Recipient false accusations were the highest in the no system condition (36.8%) and the lowest in the very strong system condition (4.6%), and they were observed more often in the weak (22.3%) than in the strong system condition (19.7%). Column (1) in Table D1 below presents results from logistic panel regressions. Recipients made significantly fewer false accusations in the very strong system than in the weak system ($X_2(1)$=15.74, p<.01), strong system ($X_2(1)$=13.10, p<.01), and no system conditions ($X_2(1)$=19.93, p<.01). Although recipients directionally made less false accusations in the strong system than in the weak system, this difference was not significant ($X_2(1)$=0.01, p=.93). Like the very strong system, both the weak system ($X_2(1)$=5.38, p=.02) and the strong system ($X_2(1)$=4.13, p=.04) were significantly better at stopping false accusations than the no system condition.

Recipient False Accusations Study 2. Recipient false accusations were more frequent in the weak system condition (16.8%) than in the very strong system condition (3.4%). The difference is statistically significant ($X_2(1)$=21.43, p<.01; see also column (2) in Table D1 below).

Summary. False accusations were strongly influenced by system strength. Stronger systems were generally better at discouraging false accusations than weaker systems. Moreover, the absence of a system lead to the highest rates of false accusations.

Table D1. Odds Ratios from Logistic Panel Regressions of System Conditions on False Accusations (Studies 1 and 2)

System	Study 1 (1)	Study 2 (2)
Strong System	1.04 ^{a} (0.49)	
Very Strong System	0.04** ^{b} (0.03)	0.11** (0.05)
No System	5.44* ^{c} (3.97)	
X_2	20.72**	21.43
N (Clusters)	608 (24)	566 (24)

Notes: The table reports odds ratios obtained from random effects logistic panel regressions. Robust standard errors are in parentheses, clustered by 24 sessions. The weak system condition is the omitted base category. An odds ratio below 1 means that the likelihood of a false accusation is smaller than in the weak system condition, an odds ratio above 1 means that it is greater. Only cases where the actor did not steal points are included. Odd ratios within column (1) that do not share a subscript are significantly different from each other (p<.05). The superscript significance markers indicate significant differences compared to the omitted weak system condition (i.e., an odds ratio significantly different than 1): $^{\dagger}p < .10$, $^{*}p < .05$, $^{**}p < .01$

Study 1 Moral Judgments and Norm Perceptions

In Study 1, we collected exploratory measures of moral judgments and norm perceptions of stealing at the end of the experiment (i.e., after round 10):

Moral judgments of stealing. Participants rated on a 5-point scale (1 = strongly disagree to 5 = strongly agree) the extent to which they thought a participant A's decision to steal (i.e., a participant A increasing his/her points by decreasing B's points) was a) fair, b) unjust, and c) right (α=.75). We reverse-coded items a) and c) to create a measure of perceived unethicality.

Norm perceptions of stealing. Participants indicated on a 5-point scale the extent to which they thought stealing decisions by As were a) not following social norms, b) normal, and c) something A should do (α=.89). We reverse-coded item a) to create a measure of perceived normativity.[2]

Table D2 below displays OLS regression results for the effects of system condition on the above measures. The weak system is the omitted base category (captured by the constant). The results indicate that system strength had an effect on moral judgments but not on norm perceptions.

Table D2. OLS Regressions of Sanctioning System on Moral Judgment and Norm Perceptions (Study 1)

System	Moral Judgment Stealing (1)	Norm of Stealing (2)
Strong System	0.15† a (0.07)	-0.00 a (0.08)
Very Strong System	0.26* a (0.10)	-0.06 a (0.12)
No System	0.02 a (0.13)	0.16* a (0.08)
Constant (No system)	3.66** (0.07)	2.68** (0.04)
R^2	.01	.01
$F(3, 23)$	2.71†	1.90
N (Clusters)	384 (24)	384 (24)

Notes: The table reports unstandardized OLS coefficients. Robust standard errors are in parentheses, clustered by 24 sessions. The dependent variables were measured on Likert scales from 1 to 5 (from "ethical" to "unethical" for moral judgment and from "not normative" to "normative" for norm perceptions). Coefficients within the same column that do not share the same subscript are significantly different from each other (p<.05). The no system condition is the base category represented by the constant. The superscript significance markers thus indicate significant differences compared to the no system condition: † $p < .10$, * $p < .05$, ** $p < .01$.

[2] Our norm perceptions measure in Study 1 thus included both prescriptive and descriptive elements of norms. Prescriptive or injunctive norms refer to beliefs about what constitutes morally approved behavior and what should be done, whereas descriptive norms describe what is typical and what most people do (Cialdini et al. 1990). Because perceptions about what is appropriate is also what is typically done (Cialdini et al. 1990), we had not differentiated between prescriptive and descriptive norms in Study 1. However, because they are conceptually distinct (Cialdini et al. 1990), we extended our norms measure in Study 2 by adding more items and separating prescriptive and descriptive norms.

Association with behavioral decisions. Logistic regression results reveal that, increased moral judgments that stealing was unethical were indeed associated with less actor stealing (*OR*=0.51, *SE*=0.07, $X_2(1)$=21.46, *p*<.01), more recipient true reporting (*OR*=1.92, *SE*=0.31, $X_2(1)$=16.07, *p*<.01), and more observer helping (*OR*=1.88, *SE*=0.35, $X_2(1)$=11.55, *p*<.01), thus providing overall correlational support for Hypothesis 4a-c. Similarly, increased perceptions that stealing was normative were significantly associated with more actor stealing (*OR*=1.76, *SE*=0.32, $X_2(1)$=9.98, *p*=.02), less recipient true reporting (*OR*=0.46, *SE*=0.10, $X_2(1)$=12.92, *p*<.01) and less observer helping (*OR*=0.50, *SE*=0.08, $X_2(1)$=16.78, *p*<.01), thus providing preliminary evidence for Hypothesis 5a-c.

APPENDIX E: ADDITIONAL RESULTS AND ANALYSES (STUDIES 1 AND 2)

Table E1. Odds Ratios from Pooled Logistic Regressions of System Conditions on Behaviors (Study 1)

System Condition	Actor Stealing (1)	Recipient Reporting (2)	Observer Helping (3)
Strong System	0.47* a,b (0.16)	1.12a (0.36)	1.30 a (0.36)
Very Strong System	0.26**a (0.09)	2.31a,b (1.19)	2.45** b (0.56)
No System	0.50**b (0.13)	4.24**b (1.81)	1.94*a,b (0.50)
Pseudo R_2	.04	.04	.02
$X_2(3)$	18.18**	13.10**	18.14**
N (Clusters)	1,280 (24)	672 (24)	472 (24)

Notes: The table reports odds ratios. Robust standard errors are in parentheses, clustered by 24 sessions. The weak system condition is the omitted base category. An odds ratio below 1 means that the likelihood of engaging in the respective behavior is smaller than in the weak system condition, an odds ratio above 1 means that it is greater. Column (2) only includes cases where the actor stole points and column (3) only includes cases where the recipient made a report, and the observer received a (possibly false) computer message that the actor had stolen points. Odds ratios within the same column that do not share a subscript are significantly different from each other (p<.05). The superscript significance markers indicate significant differences compared to the omitted weak system condition: $^{\dagger}p <$ .10, $^{*}p < .05$, $^{**}p < .01$

Figure E1. Behavioral Decisions across Rounds and Conditions (Study 1)

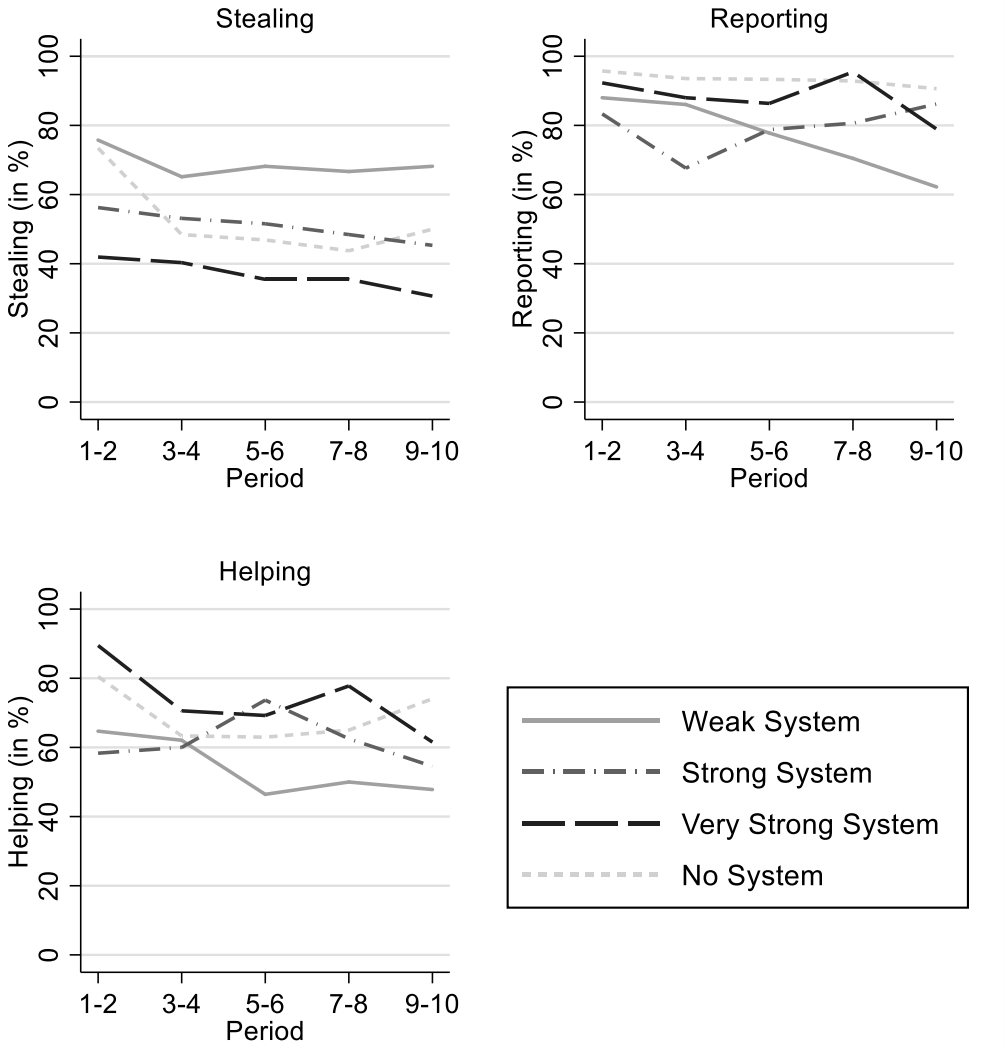


Table E2. Odds Ratios from Pooled Logistic Regressions of System Condition on Behaviors (Study 2)

System Condition	Actor Stealing (1)	Recipient Reporting (2)	Observer Helping (3)
Very Strong System	0.23** (0.13)	1.28 (0.47)	2.02* (0.63)
Pseudo R_2	.09	.00	.02
$X_2(3)$	19.64**	0.46	5.11*
N (Clusters)	1,200 (24)	634 (24)	403 (24)

Notes: The table reports odds ratios. Robust standard errors are in parentheses, clustered by 24 sessions. The weak system condition is the omitted base category. An odds ratio below 1 means that the likelihood of engaging in the respective behavior is smaller in the very strong than in the weak system condition, an odds ratio above 1 means that it is greater. Column (2) only includes cases where the actor stole points and column (3) only includes cases where the recipient made a report, and the observer received a (possibly false) computer message that the actor had stolen points. The superscript significance markers indicate significant differences compared to the omitted weak system condition: † $p < .10$, * $p < .05$, ** $p < .01$

Table E3. OLS Regressions on Perceived Unethicality and Perceived Pervasiveness Before the First Round (Study 2)

System	Perceived Unethicality (1)	Perceived Unethicality (2)	Perceived Typicality (3)	Perceived Typicality (4)
Very Strong System	0.18* (0.08)	0.23† (0.13)	-0.20* (0.08)	-0.32* (0.12)
Role B/C		0.45** (0.12)		-0.34** (0.11)
Very Strong System X Role B/C		-0.08 (0.16)		0.18 (0.16)
Constant (Weak system)	3.46** (0.06)	3.16** (0.10)	3.68** (0.05)	3.91** (0.09)
R_2	.01	.08	.02	.04
F	4.87*	9.87**	5.86*	5.88**
N	360	360	360	360

Notes: The table reports unstandardized OLS coefficients. Heteroscedasticity robust standard errors are in parentheses. The dependent variables were measured on Likert scales from 1 to 5 (from "ethical" to "unethical" and from "not typical" to "typical"). The weak system condition is the base category represented by the constant. "Role B/C" is a dummy variable that takes on value 1 if a participant was assigned the role of B/C (recipient / observer).
† $p < .10$, * $p < .05$, ** $p < .01$.

Table E4. OLS Regressions on Perceived Unethicality and Perceived Pervasiveness *After* the Final Round (Study 2)

System	Perceived Unethicality (1)	Perceived Unethicality (2)	Perceived Typicality (3)	Perceived Typicality (4)
Very Strong System	0.25** (0.08)	0.19 (0.14)	-0.94** (0.19)	-0.86** (0.20)
Role B/C		0.40** (0.12)		-0.12* (0.08)
Very Strong System X Role B/C		0.09 (0.18)		-0.12 (0.15)
Constant (Weak system)	3.56** (0.05)	3.30** (0.11)	3.75** (0.10)	3.86** (0.11)
R_2	.02	.09	.21	.22
F	9.84**	9.63**	25.01**	11.03**
N (Clusters)	360 (24)	360 (24)	360 (24)	360 (24)

Notes: The table reports unstandardized OLS coefficients. Robust standard errors are in parentheses, clustered by 24 sessions. The dependent variables were measured on Likert scales from 1 to 5 (from "ethical" to "unethical" and from "not typical" to "typical"). The weak system condition is the base category represented by the constant. "Role B/C" is a dummy variable that takes on value 1 if a participant was assigned the role of B/C (recipient / observer).
† $p < .10$, * $p < .05$, ** $p < .01$.

Table E5. Full Results for *Multiple* Mediation Analyses (Study 2)

	Indirect Effects via Perceived Unethicality		Indirect Effects via Perceived Typicality		Total Indirect Effect	
	Estimate (SE)	95% BCCI	Estimate (SE)	95% BCCI	Estimate (SE)	95% BCCI
First Round Decisions:						
Actor Stealing	-0.27 (0.19)	[-0.78, -0.01]	-0.54 (0.28)	[-1.19, -0.11]	-0.81 (0.37)	[-1.60, -0.15]
Recipient Reporting	-0.01 (0.10)	[-0.34, 0.12]	0.08 (0.15)	[-0.10, 0.59]	0.07 (0.18)	[-0.25, 0.53]
Observer Helping	0.07 (0.08)	[-0.08, 0.27]	0.01 (0.06)	[-0.08, 0.16]	0.08 (0.08)	[-0.06, 0.27]
Final Round Decisions:						
Actor Stealing	-0.08 (0.12)	[-0.45, 0.06]	-1.16 (0.39)	[-2.00, -0.59]	-1.25 (0.40)	[-2.05, -0.59]
Recipient Reporting	0.11 (0.26)	[-0.14, 1.10]	0.20 (0.30)	[-0.08, 1.21]	0.30 (0.37)	[-0.19, 1.20]
Observer Helping	-0.03 (0.05)	[-0.22, 0.03]	0.02 (0.12)	[-0.17, 0.33]	-0.01 (0.11)	[-0.24, 0.20]

Notes: N=120 for all analyses reported in the table. The table reports the indirect effects of the very strong sanctioning system via the mediators of perceived unethicality / normative inappropriateness and perceived pervasiveness of stealing on the dependent variables indicated in the first column. The top half displays results for first round decisions using the perceptions measured before the first round as mediators and the bottom half displays results for final round decisions using the perceptions measured after the final round as mediators. We implemented the *multiple* mediation procedure by Preacher & Hayes (2008) via the PROCESS macro in SPSS (Hayes, 2013). Note that PROCESS uses logistic regression to model binary dependent variables, but does not allow for the calculation of odds ratios. Bootstrapped standard errors (based on 5,000 replications) are in parentheses. PROCESS does not allow for clustering of standard errors. Reported confidence intervals are bias-corrected (BC) and thus potentially asymmetric.

Table E6. Results for *Simple* Mediation Analyses (Study 2)

	Indirect Effects via Perceived Unethicality		Indirect Effects via Perceived Pervasiveness	
	Estimate (SE)	95% BCCI	Estimate (SE)	95% BCCI
First Round Decisions:				
Actor Stealing	-0.37 (0.22)	[-0.88, 0.02]	-0.67 (0.31)	[-1.35, -0.16]
Recipient Reporting	-0.02 (0.11)	[-0.39, 0.11]	0.09 (0.14)	[-0.07, 0.57]
Observer Helping	0.07 (0.08)	[-0.05, 0.26]	0.03 (0.05)	[-0.05, 0.19]
Final Round Decisions:				
Actor Stealing	-0.21 (0.17)	[-0.62, 0.06]	-1.30 (0.38)	[-2.08, -0.68]
Recipient Reporting	0.16 (0.26)	[-0.11, 1.09]	-0.35 (0.72)	[-1.76, 1.06]
Observer Helping	-0.03 (0.05)	[-0.19, 0.02]	-0.02 (0.10)	[-0.22, 0.18]

Notes: N=120 for all analyses reported in the table. The table reports the indirect effects of the very strong sanctioning system via the mediators of perceived unethicality / normative inappropriateness and perceived pervasiveness of stealing on the dependent variables indicated in the first column. The top half displays results for first round decisions using the perceptions measured before the first round as mediators and the bottom half displays results for final round decisions using the perceptions measured after the final round as mediators. We implemented a *simple* mediation procedure via the PROCESS macro in SPSS (Hayes, 2013). Note that PROCESS uses logistic regression to model binary dependent variables, but does not allow for the calculation of odds ratios. Bootstrapped standard errors (based on 5,000 replications) are in parentheses. PROCESS does not allow for clustering of standard errors. Reported confidence intervals are bias-corrected (BC) and thus potentially asymmetric.