# Dishonest Helping and Harming After (Un)fair Treatment

Margarita Leib[1], Simone Moran[2], & Shaul Shalvi[1]

[1]University of Amsterdam; [2]Ben-Gurion University of the Negev

Supplementary Online Materials

## Pilot

**Participants and procedure**. Participants (N = 366; 75.68% female, $M_{Age}$ = 24.29, $SD_{Age}$ = 1.78) arrived at the lab in groups of 6–40 to complete an experiment in exchange for course credit and an opportunity to earn extra money. Participants were randomly assigned to one of two conditions (Allocation: dictator vs. random). Within condition, they were randomly paired with a counterpart whose identity remained anonymous throughout and after the experiment, and assigned to the role of a dictator or a recipient. In the dictator condition (N = 184; 92 dyads), dictators received 25 ILS (5 × 5 ILS coins), were asked how they wished to split the money between self and their counterpart. Dictators could choose between (1) keeping all 25 ILS to themselves and giving the counterpart 0 ILS, (2) keeping 20 ILS to themselves and giving the counterpart 5 ILS, (3) keeping 15 ILS to themselves and giving the counterpart 10 ILS, (4) keeping 10 ILS to themselves and giving the counterpart 15 ILS, (5) keeping 5 ILS to themselves and giving the counterpart 20 ILS, or (6) keeping 0 ILS to themselves and giving the counterpart 25 ILS. The dictators wrote the monetary split down on a piece of paper, and handed it to the experimenter. The experimenter subsequently placed the amount the dictators chose in an envelopes labeled “To the other person” and transferred the envelopes to the respective counterparts who was seated in another room.

In turn, recipients received their respective envelope, opened it, and learned how much money they received. Recipients knew what were the allocation options the dictators could choose from. Recipients then received ten 1 ILS coins and two envelopes: one labeled “To the study budget” the other labeled “To the other person.” They were asked to pick up each of the 10 coins, predict the outcome of a coin toss, keep the prediction in mind, toss the coin, and place the coin in one of the envelopes depending on whether their prediction was correct or not (Shalvi,

2012). If they predicted correctly, recipients were instructed to place the coin in the envelope designated for their counterpart. If they predicted incorrectly, they were instructed to place the coin in the envelope that would go back to the study budget. Participants repeated the procedure with all 10 coins. Because only the recipients knew whether their predictions were correct, they were able to lie and misreport the number of correct predictions to inflate or deflate their counterpart's pay. We assess lying at a group level by comparing the mean of reported "correct" coin-toss predictions in the different conditions with the success rate of an honest report (EV=5).

The random condition (N = 182; 91 dyads) was identical, with one exception. The dictators did not make the initial allocation decision; rather, the split of money between the dictators and the recipients was randomly determined and that was common knowledge. To keep both settings identical, the split was randomly chosen from the exact same distribution of offers made by the dictators in the dictator condition. The particular distribution of the offers was not known for the recipients. Overall, in the dictator condition, recipients have a motivation to reciprocate their counterparts' (un)fairness, whereas in the random condition they merely reacted to (un)fair treatments.

Upon completing the task, recipients completed the following scales on a 1-7 point-scale (1 = *not at all*, 7 = *definitely yes*).

**Fairness.** (1) To what extent do you feel that the amount you received is fair? (2) To what extent do you feel that the split between you and your counterpart is fair? (3) To what extent are you disappointed with the amount you received? (r) (4) To what extent do you think that this amount is not decent? (r) (5) To what extent were you happy with the amount you received? ($\alpha$ = .89)

**Gratitude**. "In the following questionnaire you will read words that describe different feelings and emotions. Please read the following words and rate the extent to which you feel at this moment: Gratitude."

**Negative feelings**. Participants were asked to rate the following items taken from the PANAS scale on a scale (Watson, Clark, & Tellegen, 1988; $\alpha = .84$): Irritable; Shame; Nervous; Jittery; Distressed; Upset; Guilty; Afraid; Hostile.

**Positive feelings**. Participants were asked to rate the following items taken from the PANAS scale (Watson, Clark, & Tellegen, 1988; $\alpha = .79$): Alert; Inspired; Determined; Interested; Excited.

**Social value orientation**. At least half an hour after the experiment was over, all participants completed the SVO scale (Decomposed Game; Messick & McClintock, 1968; Van Lange, 1999).

## Results

**Fairness.** On average, recipients received 8.20 ILS ($SD = 4.92$) from the dictators. An ANCOVA analysis with the Amount (as a continuous variable) and Allocation (dictator vs. random) predicting the level of fairness evaluated revealed an Amount × Allocation interaction, $F(1,179) = 4.21$, $p = .042$, $\eta^2 = .023$. When the monetary split was determined randomly, the higher the amount participants received, the more they evaluated it as fair, $r = .631$, $p < .001$. This correlation was even stronger for participants in the dictator condition, $r = .752$, $p < .001$.

**Recipients' behavior.** An ANCOVA analysis with the Amount (as a continuous variable) and Allocation (dictator vs. random) predicting the number of reported correct predictions (between 0 and 10) revealed no main effects for Amount, $p = .649$, and no main effect for Allocation, $p = .121$. The Amount × Allocation interaction was also not significant,

$F(1,179) = 2.97$, $p = .086$, $\eta^2 = .016$. For exploratory reasons we explored the interaction, and found that when a dictator determined the money split, a trend showed that the higher the amount recipients received, the more coin tosses they reported predicting "correctly," $r = .18$, $p = .082$. This trend did not emerge when the amount was determined randomly, $p = .426$.

We further assessed participants' dishonest helping and harming after being treated (un)fairly. To do so, we need to decide on a cutoff point from which an amount is considered fair. In the pilot, dictators could split the money in increments of 5 ILS, so implementing a cutoff point of 50% of the initial endowment (as done in Experiments 1, 2, and 3 in the manuscript) was not possible. Potential cutoff points are the mean (8.20 ILS) or median (10 ILS) amount participants received. In the dictator condition, 33 (25.86%) monetary splits were below the median (0-5 ILS), and 13 (14.13%) monetary splits were above the median (15-25 ILS). The remaining 46 (50.00%) were the median (10 ILS). Due to the small number of monetary splits above the median (n=13), when categorizing the amounts as unfair vs. fair, we opt to include the median (10 ILS) in the above the median, fair, category. We thus categorized receiving 0-5 ILS as unfair, and 10-25 ILS as fair. In both conditions, 36.07% (n=66) of the splits were thus categorized as unfair, and 63.93% (n=117) were categorized as fair.

A two-way ANOVA with 2 (Amount: unfair [0-5 ILS] vs. fair [10-25 ILS]) by 2 (Allocation: dictator vs. random) predicting the number of correct coin-toss predictions revealed no main effect for Amount, $p = .484$. Further, the main effect for Allocation, $p = .726$, and the Amount × Allocation interaction, $p = .429$, were not significant. Sensitively test revealed that the experiment was underpowered. Specifically, that with our sample size we could only detect a very large effect size ($f = 0.66$), and that we had a power of 0.10 to detect a medium effect size.

Merely for exploratory reasons we compared the number of predictions participants reported predicting "correctly" in each condition to the expected value (EV) of an honest report (EV=5). In the dictator conditions, recipients who received fair amounts reported more "correct" coin-toss predictions compared to the EV of an honest report ($M$ = 5.56; $SD$ = 1.22), $t(58) = 3.51$, $p = .001$. Recipients who received unfair amounts from a dictator, however, did not report differently than the EV of honest reports ($M$ = 5.24; $SD$ = 1.37), $t(32) = 1.01$, $p = .317$. In the random condition, participants who received fair amounts reported correctly predicting significantly more than five coin tosses ($M$ = 5.47; $SD$ = 1.39), $t(57) = 2.54$, $p = .014$, whereas those who received unfair amounts did not predict differently than the EV of honesty ($M$ = 5.48; $SD$ = 1.60), $t(32) = 1.73$, $p = .092$. When adjusting the significance level for multiple comparisons (0.05/4 = 0.012), $p < .012$ becomes significant, indicating that only after receiving a fair amount from a dictator, participants' reports are significantly higher than the EV of honesty.

**Emotions and motivations.** We ran a series of exploratory ANCOVA analyses with 2 (Allocation: dictator vs. random) by Amount (as a continuous variable) and number of correct coin toss predictions (as a continuous variable) predicting the self-reported emotion and motivation.

*Gratitude*. There was a main effect for the Amount, $F(1, 175) = 5.66$, $p = .018$, $\eta^2 = .031$. The higher the amount participants received, the more grateful they felt, $b = .355$. No other main effects or interactions were significant, $p$'s > .272.

*Positive emotions (PANAS).* No main effects or interaction were significant, $p$'s > .667.

*Negative emotions (PANAS).* There was a main effect for the Amount, $F(1, 175) = 4.01$, $p = .047$, $\eta^2 = .022$. The higher the amount participants received, the less overall negative emotions they reported, $b = -.144$. No other main effects or interactions were significant, $p$'s > .135.

### Individual differences in dishonesty

*Gender*. An ANCOVA with Allocation (dictator vs. random), Amount (as a continuous variable), and Gender (male vs. female) predicting the number of correct coin toss predictions revealed no main effect for gender, $p = .530$. Further, all the interactions with gender were not significant, $p$'s > .248.

*Age.* An ANCOVA with Allocation (dictator vs. random), Amount (as a continuous variable), and Age predicting the number of correct coin toss predictions revealed no main effect for age, $p = .932$. Further, all the interactions with age were not significant, $p$'s > .799.

*SVO*. An ANCOVA with Allocation (dictator vs. random), Amount (as a continuous variable), and number of pro social decisions in the SVO measurement (out of 9, as a continuous measure) predicting the number of correct coin toss predictions revealed no main effect for number of pro social decisions, $p = .357$. Further, all the interactions with number of pro social decisions were not significant, $p$'s > .469.

## Discussion

The pilot study assessed dishonesty at a group level only, and thus does not allow identifying for each individual whether they lied to help or harm their counterpart after being treated (un)fairly. Initial results reveal no difference in participants' behavior when the amount was determined by a dictator or randomly. Note that the pilot (1) allows assessing dishonesty only on a group level, and (2) is underpowered. Thus in the main paper we report three well powered experiments. Experiment 1 assess dishonesty at a group level, whereas in Experiment 2 and 3 we assess dishonesty on an individual level and compare to prevalence of dishonest helping and harming after being treated (un)fairly.

## Experiment 1

Upon completing the task, recipients assessed the following scales (1 = *not at all*, 7 = *definitely yes*).

**Fairness of the amount.** To what extent you feel that the amount you received from the other person is fair?

**Fairness of the counterpart.** To what extent you feel that the other person was fair?

**Generosity of the amount.** To what extent you feel that the amount you received from the other person is generous?

**Generosity of the counterpart.** To what extent you feel that the other person was generous?

**Accuracy**. While completing the task, to what extent were you motivated to be accurate?

**Gratitude.** (1) While completing the task, to what extent were you motivated to express gratitude toward the other person? (2) While completing the task, to what extent were you motivated by feeling of obligation to the other person? (3) While completing the task, to what extent were you motivated to maximize the other person's profit? (4) While completing the task, to what extent were you motivated to help the other person? (5) While completing the task, to what extent were you motivated to harm the other person? (r) ($\alpha$ = .76)

**Anger.** While completing the task, to what extent were you motivated by feeling of anger toward the other person?

**Disappointment**. While completing the task, to what extent were you motivated by feelings of disappointment toward the other person?

**Motivation to maintain fairness**. (1) While completing the task, to what extent were you motivated to maintain fairness? (2) While completing the task, to what extent were you motivated to restore justice? ($\alpha$ = .14)

**Guilt.** While completing the task, to what extent were you motivated by feelings of guilt?

**Negative feelings**. Participants were asked to rate the following items taken from the PANAS scale (Watson, Clark, & Tellegen, 1988; α = .84): Irritable; Shame; Nervous; Jittery; Distressed; Upset; Guilty; Afraid; Hostile.

**Positive feelings**. Participants were asked to rate the following items taken from the PANAS scale (Watson, Clark, & Tellegen, 1988; α = .79): Alert; Inspired; Determined; Interested; Excited.

**Social value orientation**. Participants completed the same SVO scale as in the pilot just before the main task.

## Results

**Fairness of the counterpart**. A two-way ANOVAs with the Amount (unfair [0 cents] vs. fair [15 cents]) and Framing (give-some vs. take-some), predicting the extent to which recipients evaluated their counterparts as fair, revealed a main effect of the amount. Recipients who received 15 cents evaluated their counterparts (i.e., dictators) as fairer ($M$ = 6.04, $SD$ = 1.44) than those who received 0 cents ($M$ = 2.45, $SD$ = 1.76), $F(1, 1278) = 1508.95$, $p < .001$, $\eta^2 = .541$. This was not qualified by an Amount × Framing interaction, $p = .474$.

**Generosity**. Two two-way ANOVAs with the Amount (unfair [0 cents] vs. fair [15 cents]) and Framing (give-some vs. take-some) predicting the extent to which recipients evaluated (1) the amount they received, and (2) their counterparts as generous, revealed a significant Amount × Framing interactions for the evaluation of the amount , $F(1, 1278) = 48.75$, $p < .001$, $\eta^2 = .037$, and the evaluation of the counterpart, $F(1, 1278) = 41.84$, $p < .001$, $\eta^2 = .032$. Participants who received an unfair amount (0 cents) evaluated the amount and counterpart as

generous in the give-some ($M_{amount}$ = 1.68, $SD_{amount}$ = 1.47; $M_{counterpart}$ = 1.61, $SD_{counterpart}$ = 1.42) and take-some framing ($M_{amount}$ = 1.89, $SD_{amount}$ = 1.72; $M_{counterpart}$ = 1.77, $SD_{counterpart}$ = 1.50), $p_{amount}$ = .168, $p_{counterpart}$ = .305. However, participants who received a fair amount (15 cents) evaluated both the amount and the counterpart as more generous in the give-some ($M_{amount}$ = 6.31, $SD_{amount}$ = 1.15; $M_{counterpart}$ = 6.24, $SD_{counterpart}$ = 1.28) compared to take-some faming ($M_{amount}$ = 5.21, $SD_{amount}$ = 1.67; $M_{counterpart}$ = 5.18, $SD_{counterpart}$ = 1.69), $p_{amount}$ < .001, $p_{counterpart}$ < .001.

**Accuracy.** A logistic regression with only Accuracy predicting the likelihood to report "heads" revealed no effect for motivation to be accurate, $p$ = .399. A logistic regression with Accuracy and Condition (fair vs. unfair vs. no prior treatment) predicting reporting "heads" revealed that among participants in the fair condition, the more motivated to be accurate participants reported to be, the less likely they were to report "heads", $b$ = -.145, $p$ = .004. This effect was not significant among participants in the unfair and no prior treatment conditions, $p$'s > .116.

**Emotions.** We ran a series of exploratory ANOVA analyses with 2 (Report: heads vs. tails) by 3 (Condition: unfair vs. fair vs. no prior treatment) predicting recipients' self-reported emotion and motivation.

*Gratitude.* The Report × Condition interaction was significant, $F(2, 1476) = 5.09$, $p$ = .006, $\eta^2$ = .007. Among participants who received an unfair amount, those who reported "heads" reported higher levels of gratitude than those who reported "tails", $F(1, 1476) = 20.92$, $p$ < .001, $\eta^2$ = .014. Among participants who received a fair amount, there was no difference in the level of gratitude between those who reported "heads" and those who reported "tails", $F(1, 1476) = 2.42$, $p$ = .120. Similarly, among participants who had no prior treatment, there was no difference in

the level of gratitude between those who reported "heads" and those who reported "tails", $F(1, 1476) = 0.025$, $p = .873$.

*Anger.* There was a main effect for Condition, $F(2, 1476) = 70.22$, $p < .001$, $\eta^2 = .087$. Post hoc analyses revealed that participants who received an unfair amount reported more anger ($M = 2.50$, $SD = 1.92$) than those in the fair ($M = 1.54$, $SD = 1.18$) and the no prior treatment conditions ($M = 1.43$, $SD = 1.03$), $p$'s $< .001$. There was no difference between the fair and no prior treatment conditions, $p = .926$. Lastly, the Condition × Report interaction was not significant, $p = .812$.

*Disappointment*. There was a main effect for Condition, $F(2, 1476) = 134.71$, $p < .001$, $\eta^2 = .154$. Post hoc analyses revealed that participants who received an unfair amount reported more disappointment ($M = 3.12$, $SD = 2.17$) than those in the fair ($M = 1.64$, $SD = 1.29$) and the no prior treatment conditions ($M = 1.44$, $SD = 1.01$), $p$'s $< .001$. There was no difference between the fair and no prior treatment conditions, $p = .262$. Lastly, the Condition × Report interaction was not significant, $p = .922$

**Additional scales**

*Motivation to maintain fairness.* There was a main effect for Condition, $F(2, 1476) = 7.51$, $p = .001$, $\eta^2 = .010$. Post hoc analyses revealed that participants who received an unfair amount reported lower motivation to maintain fairness ($M = 4.53$, $SD = 1.37$) than those who received a fair amount ($M = 4.84$, $SD = 1.36$), $p = .001$. No other comparisons were significant, $p$'s $> .113$. Lastly, the Condition × Report interaction was not significant, $p = .171$

*Guilt.* Neither the main effects, nor the interaction were significant, $p$'s $> .092$.

*Positive emotions (PANAS).* There was a main effect for Condition, $F(2, 1476) = 20.94$, $p < .001$, $\eta^2 = .028$. Post hoc analyses revealed that participants who received an unfair amount reported lower level of overall positive emotions ($M = 3.66$, $SD = 1.55$) than those who received a fair amount ($M = 4.22$, $SD = 1.43$) or no prior treatment ($M = 4.30$, $SD = 1.62$), $p$'s $< .001$. There was no difference between the fair and no prior treatment conditions, $p = 1.00$. Lastly, the Condition × Report interaction was not significant, $p = .740$.

*Negative emotions (PANAS).* There was a main effect for Condition, $F(2, 1476) = 14.86$, $p < .001$, $\eta^2 = .020$. Post hoc analyses revealed that participants who received an unfair amount reported higher level of overall negative emotions ($M = 1.85$, $SD = 1.13$) than those who received a fair amount ($M = 1.53$, $SD = 1.00$) or no prior treatment ($M = 1.47$, $SD = 0.85$), $p$'s $< .001$. There was no difference between the fair and no prior treatment conditions, $p = 1.00$. Lastly, the Condition × Report interaction was not significant, $p = .990$.

| | n | Motivation to maintain fairness | Guilt | Positive emotions | Negative emotions |
|---|---|---|---|---|---|
| **Unfair amount (0 cents)** | | | | | |
| Reporting "heads" | 200 | 4.59 (1.34) | 1.96 (1.52) | 3.69 (1.60) | 1.86 (1.12) |
| Reporting "tails" | 180 | 4.46 (1.40) | 1.68 (1.35) | 3.62 (1.51) | 1.84 (1.13) |
| **Fair amount (15 cents)** | | | | | |
| Reporting "heads" | 560 | 4.82 (1.36) | 1.97 (1.53) | 4.22 (1.45) | 1.55 (1.03) |
| Reporting "tails" | 342 | 4.88 (1.37) | 2.08 (1.57) | 4.21 (1.40) | 1.51 (0.95) |
| **No prior treatment** | | | | | |
| Reporting "heads" | 126 | 4.66 (1.49) | 1.85 (1.43) | 4.25 (1.59) | 1.48 (0.91) |
| Reporting "tails" | 74 | 4.98 (1.35) | 2.01 (1.59) | 4.38 (1.68) | 1.44 (0.74) |

**Table S1.** Means (SD) of the level of motivation to maintain fairness, guilt, positive and negative emotions per condition (unfair vs. fair vs. no prior treatment) and whether participants reported the beneficial outcome for the counterpart (heads) or not (tails). Significance level: *** $p < .001$ for the difference from the cell above. Adjusting significance level for all the measures we collected (7 in total), the new significance level is 0.05/7 = 0.007. $p < .007$ will be considered significant, thus all comparisons marketed as *** remain significant.

### Individual differences in dishonesty

*Gender*. A chi-square analysis revealed no differences between males and females in reporting the beneficial outcome, "heads", $\chi^2(1) = 0.32$, $p = .858$. In particular, 59.45% of the males and 59.94% of the females reported "heads." Further, there was no difference between males and females in reporting "heads" in the unfair, fair, and no prior treatment conditions, $p$'s > . 114.

*Age.* A logistic regression with Age and Condition predicting reporting "heads" revealed that age did not predicted the likelihood of reporting "heads", $p = .637$. Further, Age × Condition interactions were not significant, $p$'s > .726.

*SVO*. A logistic regression with only the number of pro-social decisions (out of 9) predicting the likelihood to report "heads" revealed no effect for number of pro-social decisions, $p = .094$. A logistic regression with number of pro-social decisions and Condition predicting reporting "heads" revealed that among participants in the fair condition, the more pro social decisions participants made in the SVO measure, the more likely they were to report "heads", $b = .038$, $p = .027$. This effect was not significant among participants in the unfair and no prior treatment conditions, $p$'s > .174.

## Experiment 2

After the task, recipients evaluated the following scales (1 = *not at all*, 7 = *definitely yes*):

**Fairness of amount**: To what extent you feel that the amount you received from the other person is fair?

**Fairness of counterpart:** To what extent you feel that the other person was fair?

**Accuracy**. While completing the task, to what extent were you motivated to be accurate in the task?

**Gratitude.** (1) While completing the task, to what extent were you motivated to express gratitude toward the other person? (2) While completing the task, to what extent were you motivated by feeling of obligation to the other person? (3) While completing the task, to what extent were you motivated to maximize the other person's profit? (4) While completing the task, to what extent were you motivated to help the other person? (5) While completing the task, to what extent were you motivated to harm the other person? (r) ($\alpha = .86$)

**Anger.** While completing the task, to what extent were you motivated by feeling of anger toward the other person?

**Disappointment**. While completing the task, to what extent were you motivated by feelings of disappointment toward the other person?

**Motivation to maintain fairness**. (1) While completing the task, to what extent were you motivated to maintain fairness? (2) While completing the task, to what extent were you motivated to restore justice? ($\alpha = .44$)

**Guilt.** While completing the task, to what extent were you motivated by feelings of guilt?

**Social value orientation**. Participants completed the same SVO scale as in previous experiments when they signed up to the experiment (1–2 day before the experiment in the lab).

**Normative believe.** We assess whether recipients' behavior is driven by the amount they believe is normative to give in this setting. To do so, after completing the task, participants reported how much (out of 20 ILS) they would have given if they were in the role of the dictator.

## Results

**Fairness of the counterpart.** The higher the amount participants received, the fairer they evaluated their counterpart to be, $r = .68$, $p < .001$.

### Factors of the ambiguous die paradigm

In the following set of analyses we assessed whether the type of misreports (helping vs. harming) were affected by additional factors of the task. We thus focused on the trials in which participants misreported the outcome, and assessed whether the amount they received and additional factors of the task affect the type of misreports.

*Gap between target and value next to the target.* We assessed whether the likelihood to misreport an outcome that helped versus harmed the dictator was affected by the gap between the target and the value near the target. We thus calculated an absolute gap between the value next to the target and the correct outcome, resulting in coding of '1' for trails in which the value next to the target was 2 or 4, and coding of '2' for trails in which the value next to the target was 1 or 5. A generalized linear mixed model predicting the likelihood of misreporting a helpful (vs. harmful) value with the Amount participants received (as a continuous variable) and the Gap (1 vs. 2) revealed that the two way Amount × Gap interaction was not significant, $p = .286$. Further, the main effect for Gap was also not significant, $p = .085$. The main effect for Amount was significant, $b = 0.318$, $p = .003$, indicating that the higher the amount participants received the more likely they were to make helpful, compared to harmful misreports.

*Target location*. A generalized linear mixed model predicting the likelihood of misreporting a helpful (vs. harmful) value with the Amount participants received (as a continuous variable) and the Target Location revealed that the Amount × Target Location interaction was not significant, $p = .942$. Further, the main effect for Target Location was also not significant, $p = .721$. The main effect for Amount was significant, $b = 0.263$, $p = .018$,

indicating that the higher the amount participants received the more likely they were to make helpful, compared to harmful misreports.

*Fixation cross location.* A generalized linear mixed model predicting the likelihood of misreporting a helpful (vs. harmful) value with the Amount participants received (as a continuous variable) and the Fixation Cross Location revealed that the Amount × Fixation Cross Location interaction was not significant, $p = .545$. Further, the main effect for Fixation Cross Location was also not significant, $p = .902$. The main effect for Amount was significant, $b = 0.306$, $p = .010$, indicating that the higher the amount participants received the more likely they were to make helpful, compared to harmful misreports.

**Robustness checks**

*Trial number.* Figure S1 present the mean report for every trial number separated for the participants who received an unfair (0-8 ILS; dashed line) and fair (10-20 ILS; solid line) amounts. As can be seen, participants who received a fair amount reported systematically higher values than participants who received unfair amounts. Further, there was no clear pattern of reports across trial number. To investigate this further, we conducted a linear mixed model regression predicting the value recipients reported from the Amount they received (as a continuous variable), the Trial Number (as a continuous variable). Results reveal a main effect for the Amount, $b = .052$, $t(76.8) = 2.10$, $p = .038$. There was no main effect for the Trial Number, $p = .528$. Further, the Amount × Trial Number interaction was not significant, $p = .153$.

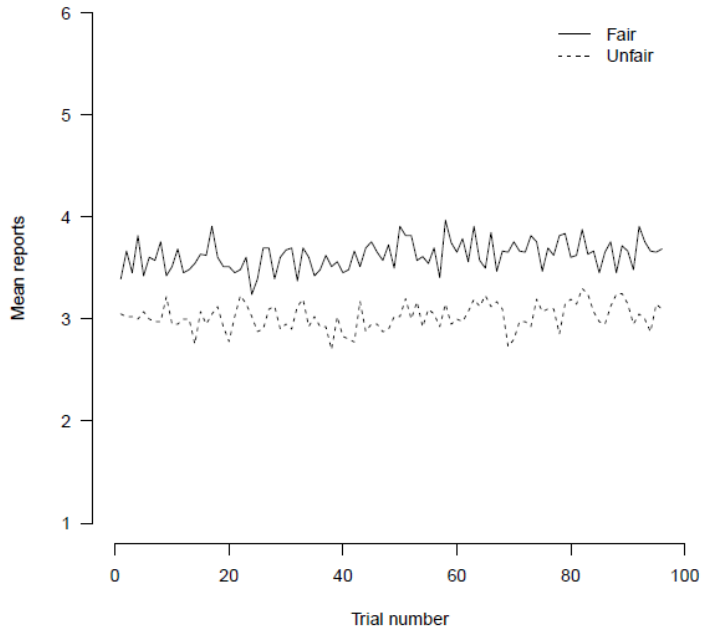


*Figure S1.* The mean report of participants who received a fair amount (10 ILS or more) and an unfair amount (less than 10 ILS), as a function of trial number.

*Amount received.* Figure S2 plots the behavioral types as a function of the amount received as a continuous measure. In the figure, every participant is represented by one diamond. Green indicates that the participant was classified as a dishonest helper, red indicates that the participant was classified as a dishonest harmer, and gray indicates that the participants was classified as inconsistent. The figure shows that even among participants who received 0, 2, 4, and 6 ILS, some were dishonest helpers. That is, dishonestly helping the counterpart after receiving a rather unfair amount was not restricted to the higher values that were classified as unfair.

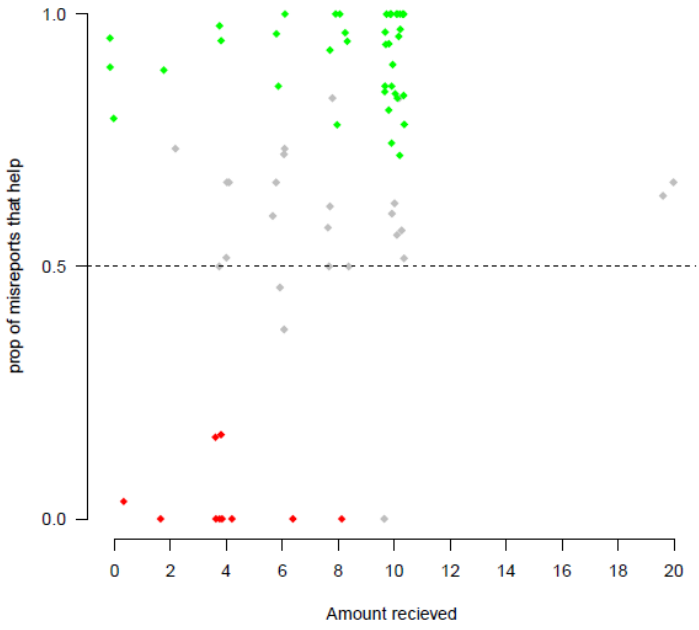


*Figure S2.* The proportion of misreports the helped the dictator out of all misreports as a function of the amount participants received (as a continuous measure). Green diamond indicates participants who was classified as a "dishonest helper", red diamond indicates participants who was classified as a "dishonest harmer", and gray indicates participants who were classified as "inconsistent".

*Normative believe.* The amount recipients thought is appropriate to give in the task did not affect their reports. A linear regression with the amount participants received and the amount participants would have given had they were in the role of the dictator, predicting helping misreports (out of all misreports) revealed a significant effect for the amount received, $b = 0.058$, $t(70) = 2.16$, $p = .034$. The main effect for the amount participants would have given as dictators and the interaction between how much they received and how much they would have given were not significant, $p = .427$ and $p = .229$, respectively.

**Different fairness classifications**

We assess whether our results robust to various fairness classifications.

*Median and mean split of the amount and 30% of the initial endowment*. The average amount received was 7.56 ILS, and the median was 8 ILS. Thus, a median or mean split will dictate that 0-6 ILS is an unfair amount, whereas 8-20 ILS is a fair amount. This will also be the classifications if all amounts above 30% of the initial endowment will be classified as fair, and 30% or less will be classified as unfair. A chi square analysis employing this cutoff (unfair: 0-6 ILS, fair: 8-20 ILS) revealed that the proportion of dishonest helpers among participants who received a fair amount (66.66%) was higher than the proportion of dishonest harmers among participants who received an unfair amount (31.03%), $\chi^2(1) = 8.98$, $p = .005$, Cramer's $V = .348$.

*Median split of subjective evaluation of the fairness of the amount.* Participants' median evaluation of fairness of the amount was 5 on a 1-7 scale. Thus evaluations of 1-5 were classified as unfair, and evaluations of 6-7 were classified as fair. A chi square analysis employing the median evaluation of fairness of the amount as a cutoff revealed that the proportion of dishonest helpers among participants who received a fair amount (77.50%) was higher than the proportion of dishonest harmers among participants who received an unfair amount (26.31%), $\chi^2(1) = 17.52$, $p < .001$, Cramer's $V = .487$.

*Median split of evaluation of the fairness of the counterpart.* Participants' median evaluation of the fairness of the counterpart was 6 on a 1-7 scale, thus evaluations of 1-6 were classified as unfair, and evaluations of 7 was classified as fair. A chi square analysis employing the median evaluation of fairness of the counterpart as a cutoff revealed that the proportion of dishonest helpers among participants who received a fair amount (69.44%) was higher than the proportion of dishonest harmers among participants who received an unfair amount (26.31%), $\chi^2(1) = 13.79$, $p < .001$, Cramer's $V = .432$.

*Accuracy.* Participants who were classified as dishonest helpers or harmers reported that they were less motivated to be accurate in the task ($M = 4.57$, $SD = 2.33$) than those who were classified as inconsistent ($M = 6.21$, $SD = 1.18$), $F(1, 71) = 10.42$, $p = .002$, $\eta^2 = .128$, indicating that participants' misreported were rather intentional. In line with participants' misreports being rather intentional, adding the motivation to be accurate in the task into the model attenuated the effect of the amount participants received on their misreports. A linear regression with the Amount participants received and Accuracy, predicting the proportion of misreports that help out of all misreport revealed that the effect of Amount was no longer significant, $b = 0.019$, $t(70) = 1.92$, $p = .058$. Accuracy, predicted the proportion of misreported that help out of all misreports, $B = 0.038$, $t(70) = 2.22$, $p = .029$.

**Emotions**. We ran a series of exploratory ANOVA analyses with 2 (Behavior: dishonest helping after fair treatment/dishonest harming after unfair treatment vs. not) by 2 (Amount: unfair [0-8 ILS] vs. fair [10-20 ILS]) predicting recipients' self-reported emotion and motivation.

*Gratitude.* The Behavior × Amount interaction was significant, $F(1, 70) = 9.86$, $p = .002$, $\eta^2 = .124$. Among participants who received unfair amounts, those who engaged in dishonest harming reported lower levels of gratitude than those who did not engage in dishonest harming, $F(1, 70) = 31.90$, $p < .001$, $\eta^2 = .313$. Among participants who received fair amounts, there was no difference in the level of gratitude between those who engaged in dishonest helping and those who did not engaged in it, $F(1, 70) = 0.93$, $p = .338$.

*Anger.* The Behavior × Amount interaction was significant, $F(1, 70) = 49.68$, $p < .001$, $\eta^2 = .415$. Among participants who received unfair amounts, those who engaged in dishonest harming reported higher levels of anger than those who did not engage in dishonest harming, $F(1, 70) = 115.82$, $p < .001$, $\eta^2 = .623$. Among participants who received fair amounts, there was

no difference in the level of anger between those who engaged in dishonest helping and those who did not engaged in it, $F(1, 70) = 0.14$, $p = .701$.

*Disappointment*. The Behavior × Amount interaction was significant, $F(1, 70) = 34.47$, $p < .001$, $\eta^2 = .330$. Among participants who received unfair amounts, those who engaged in dishonest harming reported higher levels of disappointment than those who did not engage in dishonest harming, $F(1, 70) = 77.27$, $p < .001$, $\eta^2 = .525$. Among participants who received fair amounts, there was no difference in the level of anger between those who engaged in dishonest helping and those who did not, $F(1, 70) = 0.02$, $p = .874$.

**Additional scales**

*Motivation to maintain fairness.* There was a main effect for Amount, $F(1, 70) = 6.27$, $p = .015$, $\eta^2 = .082$. Participants who received an unfair amount reported lower motivation to maintain fairness ($M = 4.52$, $SD = 1.55$) than those who received a fair amount ($M = 5.19$, $SD = 1.81$). The Behavior × Amount interaction was not significant, $p = .094$.

*Guilt.* Neither the main effects, nor the interaction were significant, $p$'s > .109.

| | n | Motivation to maintain fairness | Guilt |
|---|---|---|---|
| **Unfair amounts (0-8 ILS)** | | | |
| Dishonest harming | 10 | 4.40 (0.87) | 2.00 (1.11) |
| No dishonest harming | 31 | 4.56 (1.73) | 1.35 (0.66) |
| **Fair amounts (10-20 ILS)** | | | |
| Dishonest helping | 24 | 4.75 (1.90) | 1.66 (1.04) |
| No dishonest helping | 9 | 6.38 (0.78)* | 2.00 (2.12) |

**Table S2.** Means (SD) of the level of motivation to maintain fairness and guilt, per amount received (unfair 0-8 ILS; fair: 10-20 ILS) and whether participants did or did not engage in dishonest harming/helping after (un)fair treatment. Significance level: * $p < .05$; ** $p < .01$, *** $p < .001$ for the difference from the cell above. Adjusting significance level for all the measures we collected (5 in total), the new significance level is 0.05/5 = 0.01. $p < .01$ will be considered significant, thus all comparisons marked as ** and *** remain significant.

### Individual differences in dishonesty

*Gender*. A regression analyses with Amount and Gender predicting the proportion of misreports that help out of all misreports revealed that the higher the amount recipients received, the higher the proportion of misreport that help was, $B = 0.25$, $p = .015$. Gender had no significant effect, $p = .624$.

*Age.* A regression analyses with Amount and Age predicting the proportion of misreports that help out of all misreports revealed that the higher the amount received, the higher the proportion of misreport that help was, $B = 0.24$, $p = .016$. Age had no significant effect, $p = .343$.

*SVO*. A regression analyses with Amount and the Number of pro social decisions in the SVO task (out of 9) predicting the proportion of misreports that help out of all misreports revealed that the higher the amount received, the higher the proportion of misreport that help was, $B = 0.28$, $p = .004$. The number of pro social decisions had no significant effect, $p = .659$.

## Experiment 3

After the task, recipients evaluated the following scales (1 = *not at all*, 7 = *definitely yes*):

**Fairness of amount**: To what extent you feel that the amount you received from the other person is fair? For receivers in the random condition, this item read: To what extent you feel that the amount you received is fair?

**Fairness of counterpart:** To what extent you feel that the other person was fair?

**Generosity of amount**: To what extent you feel that the amount you received from the other person is generous? For receivers in the random condition, this item read: To what extent you feel that the amount you received is generous?

**Generosity of counterpart**: To what extent you feel that the other person was generous?

**Accuracy**. While completing the task, to what extent were you motivated to be accurate in the task?

**Gratitude.** (1) While completing the task, to what extent were you motivated by feelings of gratitude toward the other person? (2) While completing the task, to what extent were you motivated by feelings of obligation toward the other person? (3) While completing the task, to what extent were you motivated to maximize the other person's profit? (4) While completing the task, to what extent were you motivated to help the other person? (5) While completing the task, to what extent were you motivated to harm the other person? (r) ($\alpha = .76$)

**Anger.** While completing the task, to what extent were you motivated by feeling of anger toward the other person?

**Disappointment**. While completing the task, to what extent were you motivated by feelings of disappointment toward the other person?

**Motivation to maintain fairness**. (1) While completing the task, to what extent were you motivated to maintain fairness? (2) While completing the task, to what extent were you motivated to restore justice? ($\alpha = .35$)

**Guilt.** While completing the task, to what extent were you motivated by feelings of guilt?

## Results

**Fairness of counterpart.** When evaluating the fairness of the counterpart, the Amount × Allocation interaction was significant, $F(1,193) = 26.38$, $p < .001$, $\eta^2 = .120$. Participants in the dictator condition evaluated their counterpart as fairer when they received 10 ILS ($M = 6.55$, $SD = 1.15$) compared to 2 ILS ($M = 2.24$, $SD = 1.83$), $F(1,193) = 135.66$, $p < .001$, $\eta^2 = .413$. This

gap was attenuated in the random condition, but remained significant ($M_{fair}$ = 5.49, $SD_{fair}$ = 1.83; $M_{unfair}$ = 3.87, $SD_{unfair}$ = 2.07), $F(1,193) = 19.16$, $p < .001$, $\eta^2 = .090$.

**Generosity**. A 2 (Amount: unfair [2 ILS] vs. fair [10 ILS]) by 2 (Allocation: dictator vs. random) predicting the extent to which participants' evaluated the amount, and the counterpart as generous revealed an Amount × Allocation interaction for amount, $F(1,195) = 49.35$, $p < .001$, $\eta^2 = .202$, and counterpart, $F(1,194) = 39.02$, $p < .001$, $\eta^2 = .167$. Participants in the dictator condition evaluated their counterpart and the amount they received as more generous when they received 10 ILS ($M_{counterpart}$ = 6.20, $SD_{counterpart}$ = 1.38; $M_{amount}$ = 6.28, $SD_{amount}$ = 1.10) than 2 ILS ($M_{counterpart}$ = 1.41, $SD_{counterpart}$ = 1.01; $M_{amount}$ = 1.14, $SD_{amount}$ = 0.50), $F's > 210.02$, $p's < .001$, $\eta^2$'s > .520. This gap was attenuated in the random condition, but remained significant. Participants in the random condition evaluated their counterpart and the amount they received as more generous when they received 10 ILS ($M_{counterpart}$ = 4.60, $SD_{counterpart}$ = 2.03; $M_{amount}$ = 4.13, $SD_{amount}$ = 1.96) than 2 ILS ($M_{counterpart}$ = 2.73, $SD_{counterpart}$ = 1.90; $M_{amount}$ = 1.28, $SD_{amount}$ = 0.78), $F's > 32.00$, $p's < .001$, $\eta^2$'s > .142.

### Factors of the ambiguous die paradigm

In the following set of analyses we assessed whether the type of misreports (helping vs. harming) was affected by additional factors of the task. We thus focused on the trials in which participants misreported the outcome, and assessed whether the amount they received and additional factors of the task affect the type of misreports.

*Gap between target and value next to the target.* We assessed whether the likelihood to misreport an outcome that helps vs. harms the dictator is affected by the gap between the target and the value near the target. As in Experiment 2 we calculated the absolute gap between the value next to the target and the correct outcome. A generalized linear mixed model predicting the

likelihood of misreporting a helpful (vs. harmful) value with the Amount participants received (2 vs. 10 ILS) and the Gap revealed no Amount × Gap interaction, $p = .606$. The main effect for Gap was also not significant, $p = .236$.

*Target location*. A generalized linear mixed model predicting the likelihood of misreporting a helpful (vs. harmful) value with the amount participants received (2 vs. 10 ILS) and the Target Location revealed that the Amount × Target Location interaction was not significant, $p = .519$. Further, the main effect for Target Location was also not significant, $p = .806$.

*Fixation cross location.* A generalized linear mixed model predicting the likelihood of misreporting a helpful (vs. harmful) value with the amount participants received (2 vs. 10 ILS) and the Fixation Cross Location revealed that the Amount × Fixation Cross Location interaction was not significant, $p = .612$. Further, the main effect for Fixation Cross Location was also not significant, $p = .633$.

**Robustness checks**

*Trial number and the effect of the source of the money in the beginning of the task.* Figure S3 and S4 present the mean report for every trial number, separately for the amount participants received and the source of the money (dictator vs. random). Figure S3 presents the mean report per trial for participants in the dictator condition (solid line) and random condition (dashed line) who received an unfair allocation (2 ILS out of 20 ILS). Figure S4 presents the mean report per trial for participants in the dictator condition (solid line) and random condition (dashed line) who received a fair allocation (10 ILS out of 20 ILS). As can be seen, there is no obvious, non-linear pattern in recipients reports. Further, there is no effect on the first rounds of the task that disappears over time.

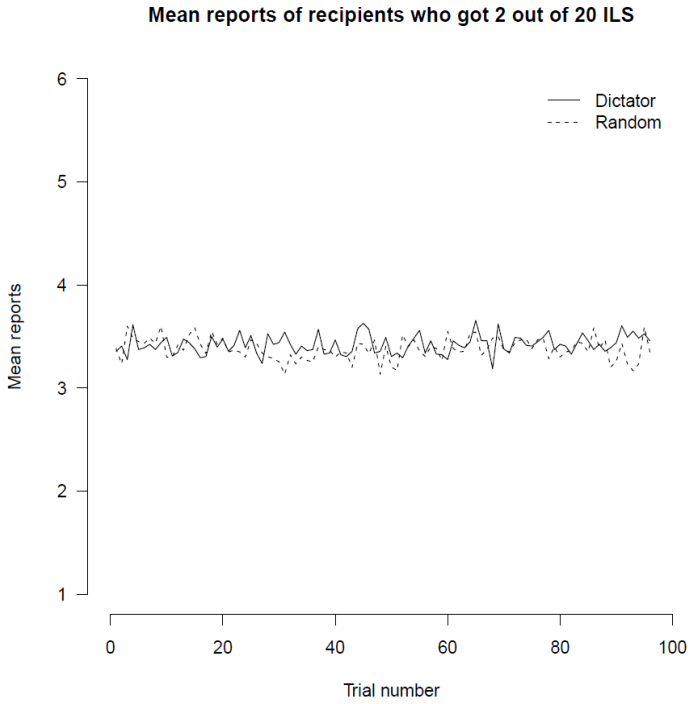


*Figure S3.* Mean report of participants who received an unfair amount (2 ILS out of 20 ILS) in the dictator (solid line) and random (dashed line) conditions, as a function of trial number.

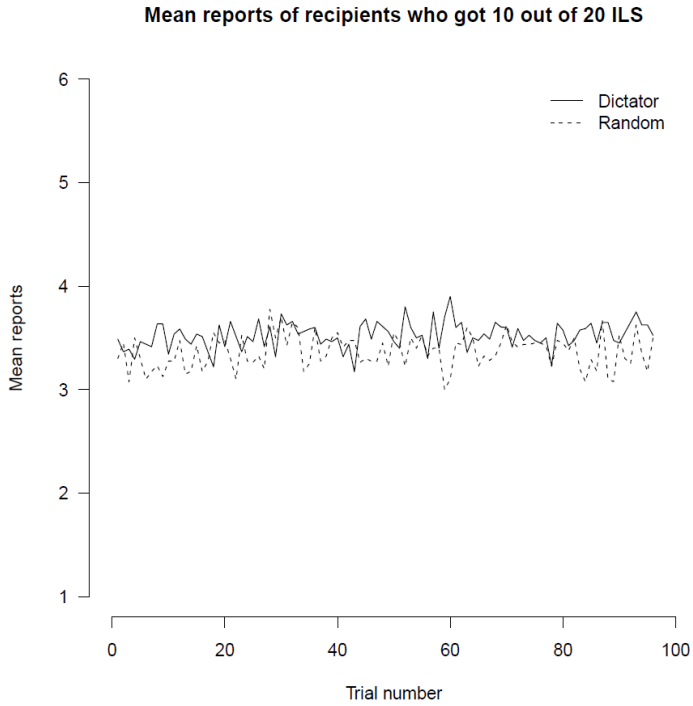


*Figure S4.* Mean report of participants who received an fair amount (10 ILS out of 20 ILS) in the dictator (solid line) and random (dashed line) conditions, as a function of trial number.

To further explore the possibility that the source of the money (dictator vs. random) affected participants' behavior only in the beginning of the task, we ran several exploratory analyses. Specifically, we restricted our analysis to the first trial of the task (trial number = 1) and the first trial after participants read a reminder of the payoff scheme (trial number = 49). We then predicted the value recipients reported from the amount they reviewed (2 ILS vs. 10 ILS) and the Allocation condition (dictator vs. random). A mixed model analysis revealed no main effect for Amount, $p = .576$, no main effect for Allocation, $p = .744$, and no Amount × Allocation interaction, $p = .794$.

Restricting our analysis to the first five trials ( $1 \leq$ trial number $\leq 5$ and $49 \leq$ trial number $\leq 54$) lead to the same results, with no main effect for Amount, $p = .859$, no main effect for Allocation, $p = .862$, and no Amount × Allocation interaction, $p = .616$. Lastly, restricting our analysis to the first ten trials ($1 \leq$ trials number $\leq 10$ and $49 \leq$ trial number $\leq 59$) lead to the same results, with no main effect for Amount, $p = .670$, no main effect for Allocation, $p = .884$, and no Amount × Allocation interaction, $p = .542$. We thus conclude that participants did not react differently to the amount they received when they source is a dictator vs. random, even in the beginning of the task.

*Accuracy.* Participants who were classified as dishonest helpers or harmers reported that they were less motivated to be accurate in the task ($M = 4.70$, $SD = 2.14$) than those who were classified as inconsistent ($M = 6.00$, $SD = 1.53$), $F(1, 195) = 21.46$, $p < .001$, $\eta^2 = .009$, indicating that participants' misreports were rather intentional. In line with participants' misreports being rather intentional, adding the motivation to be accurate in the task into the model attenuated the effect of the amount participants received on their misreports. An ANOVA with the Amount received (2 ILS vs. 10 ILS) and Accuracy (as a continues measure), predicting

the proportion of misreports that help out of all misreport revealed that the effect of Amount received was no longer significant, $F(1, 193) = 3.79$, $p = .052$. Accuracy, predicted the proportion of misreported that help out of all misreports $F(1, 193) = 4.54$, $p = .034$.

**Emotions.** We ran a series of ANOVA analyses with 2 (Behavior: dishonest helping /harming vs. not) by 2 (Amount: fair [10 ILS] vs. unfair [2 ILS]) by 2 (Allocation: dictator vs. random) predicting recipients' self-reported emotions and motivations.

*Gratitude.* The Allocation × Amount × Behavior three way interactions was not significant, $p = .524$. The Behavior ×Amount interaction was significant, $F(1, 191) = 8.64$, $p = .004$, $\eta^2 = .015$. Among participants who received an unfair amount (2 ILS), those who engaged in dishonest harming reported lower levels of gratitude than those who did not engage in dishonest harming, $F(1, 191) = 21.38$, $p < .001$, $\eta^2 = .101$. Among participants who received a fair amount (10 ILS), there was no difference in the level of gratitude between those who engaged in dishonest helping and those who did not, $F(1, 191) = 1.11$, $p = .293$.

*Anger.* The Allocation × Amount × Behavior three way interactions was not significant, $p = .090$. The Behavior × Amount interaction was significant, $F(1, 191) = 40.35$, $p < .001$, $\eta^2 = .177$. Among participants who received an unfair amount (2 ILS), those who engaged in dishonest harming reported higher levels of anger than those who did not engage in dishonest harming, $F(1, 187) = 46.55$, $p < .001$, $\eta^2 = .199$. Among participants who received a fair amount (10 ILS), there was no difference in the level of anger between those who engaged in dishonest helping and those who did not, $F(1, 187) = 2.33$, $p = .128$.

*Disappointment*. The Allocation × Amount × Behavior three way interactions was not significant, $p = .097$. The Behavior × Amount interaction was significant, $F(1, 190) = 32.42$, $p < .001$, $\eta^2 = .146$. Among participants who received an unfair amount (2 ILS), those who engaged

in dishonest harming reported higher levels of disappointment than those who did not engage in dishonest harming, $F(1, 190) = 41.95$, $p < .001$, $\eta^2 = .181$. Among participants who received a fair amount (10 ILS), there was no difference in the level of disappointment between those who engaged in dishonest helping and those who did not, $F(1, 190) = .796$, $p = .373$.

**Additional scales**

*Motivation to maintain fairness.* The Allocation × Amount × Behavior three way interactions was not significant, $p = .595$. Further, the Behavior × Amount interaction was not significant, $p = .070$. The main effect for Behavior was not significant, $p = .326$.

*Guilt.* The Allocation × Amount × Behavior three way interactions was not significant, $p = .628$. Further, the Behavior × Amount interaction was not significant, $p = .470$. The main effect for Behavior was not significant, $p = .569$.

| | n | Motivation to maintain fairness | Guilt |
|---|---|---|---|
| **Unfair amounts (2 ILS)** | | | |
| Dishonest harming | 12 | 4.87 (1.17) | 1.50 (1.00) |
| No dishonest harming | 107 | 4.00 (1.67)$^{+}$ | 1.24 (0.78) |
| **Fair amounts (10 ILS)** | | | |
| Dishonest helping | 49 | 4.93 (1.56) | 1.44 (1.35) |
| No dishonest helping | 31 | 5.17 (1.54) | 1.55 (1.33) |

**Table S3.** Means (SD) of the level of motivation to maintain fairness and guilt per amount received (unfair: 2 ILS; fair: 10 ILS) and whether participants did or did not engage in dishonest harming/helping after (un)fair amount. Since the three-way interaction with allocation (random vs. dictator) was not significant, the means reported here are collapsed across the allocation condition. Significance level: $^{+}p < .10$; * $p < .05$; ** $p < .01$, *** $p < .001$ for the difference from the cell above. Adjusting significance level for all the measures we collected (5 in total), the new significance level is 0.05/5 = 0.01. $p < .01$ will be considered significant, thus all comparisons marked as ** and *** remain significant.

**Individual differences in dishonesty.**

*Gender*. An ANOVA with Amount and Gender predicting the proportion of misreports that help out of all misreports revealed that the amount received (2 vs. 10 ILS) predicted the

proportion of helping reports, $F(1, 196) = 4.00$, $p = .047$, $\eta^2 = .020$. Gender had no significant effect on the proportion of helping misreport, $p = .776$.

*Age*. An ANCOVA with Amount and Age predicting the proportion of misreports that help out of all misreports revealed that the Amount received (2 vs. 10 ILS) no longer significantly predicted the proportion of helping reports, $p = .057$. Age had no significant effect on the proportion of helping misreport, $p = .896$.

**Deviations from pre-registration**

To ensure transparency, we list all deviations from the pre-registration and the amendment to the pre-registration that we uploaded on OSF below:

1. In the pre-registration we specified that we will calculate the gap between misreports that help and misreports that harm the counterpart and assess how the gap is affected by the amount received and allocation condition. After reflecting on the editor's and reviewers' comments, we ended up calculating and reporting the proportion of misreports that help out of the total number of misreports as our DV.

2. In the pre-registration we specified that we will focus on three 'types' (dishonest helpers after a fair amount, dishonest helpers after an unfair amount, and dishonest harmers after unfair amount). However, we realized that such approach will exclude information about participants who engaged in other behaviors (e.g., were inconsistent, dishonestly harmed after fair amounts). Thus, in order to provide full information about participants' behavior and emotions associated with those behaviors we analyzed the data assessing how different reactions (dishonestly help after fair amount/dishonestly harm after unfair amount vs. not) following the same treatment (unfair vs. fair) are associated with various emotions and motivations.

3. In the pre-registration we used labels such as “dishonest positive reciprocator” and “dishonest negative reciprocator”. Since results of Experiment 3 show that participants’ behavior was not driven by the motivation to reciprocate, such labels are suboptimal and inaccurate. In the paper we changed the labels to “dishonestly help” and “dishonestly harm” after being treated (un)fairly. Accordingly, the predictions reported in the paper are not worded similarly to the pre-registration.

**References**

Messick, D. M., & McClintock, C. G. (1968). Motivational basis of choice in experimental games. *Journal of Experimental Social Psychology*, *4*, 1-25.

Shalvi, S. (2012). Dishonestly increasing the likelihood of winning. *Judgment and Decision Making, 7*(3), 292.

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*, 337-349.

Watson, D., Clark, L. A., & Tellegen, A., (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54,* 1063-1070.