(Dis)honesty in individual and collaborative settings

A behavioral ethics approach

Leib, M.

Publication date
2021

Document Version
Other version

License
Other

Chapter 3

Dishonest helping and harming after (un)fair treatment

This chapter was published as Leib, M., Moran, S., & Shalvi, S. (2019). Dishonest helping and harming after (un) fair treatment. *Judgment and Decision Making, 14*(4), 423-439. http://journal.sjdm.org/19/190419/jdm190419.html

People experience fair and unfair treatment almost on a daily basis. As a response, people may "balance the scale", even by breaking ethical rules and lying. For instance, after being denied an expected promotion, an employee may claim higher mileage than he or she actually used on a business expense form. Conversely, after getting an exceptional bonus, an employee might lie to cover for his boss's mistake. Here, we test the prominence of dishonest behavior aimed at helping and harming others after (un)fair experiences. Specifically, we ask: what triggers higher levels of dishonesty, experiencing a fair or an unfair treatment[3]? Is dishonest harming after unfair treatment more or less common than dishonest helping after fair treatment? And does the cause of (un)fairness– whether it is intentional or not – matter?

While people value their honesty, and do not lie to a full extent (Abeler, Nosenzo, & Raymond, 2019; Mazar, Amir, & Ariely, 2008; Shalvi, Dana, Handgraaf, & De Dreu, 2011), previous work has shown that they lie to affect their own, as well as others' outcomes. Indeed, people lie to benefit themselves (Pulfrey & Butera, 2013; Shalvi, Gino, Barkan, & Ayal, 2015; Van Yperen, Hamstra, & van der Klauw, 2011), and even more so to benefit others as well (Cohen, Gunia, Kim-Jun, & Murnighan, 2009; Conrads, Irlenbusch, Rilke, & Walkowitz, 2013; Weisel & Shalvi, 2015; Wiltermuth, 2011). For example, Gino, Ayal, and Ariely (2013) found that the larger the group that can benefit from lying, the more likely are people to lie.

People also lie to harm others, especially when it is compatible with their goals. For instance, people lie to harm others when it is financially beneficial for them to do so (Erat & Gneezy, 2012; Gneezy, 2005), when it allows them to restore equity between unequal parties (Gino & Pierce, 2009; 2010a, 2010b; Moran & Schweitzer, 2008), and as a mean to financially compensated themselves after experiencing defection in a prisoners dilemma (Ellingsen, Johannesson, Lilja, & Zetterqvist, 2009). Further, when people's behavior does not affect their own outcome, they lie to harm charities they dislike (i.e., vindictive cheating; Ayal, 2015). Although there is evidence for dishonest helping and harming, one key question remains open: which treatment, fair or unfair, elicits stronger dishonest reactions? And does dishonest reactions differ when the preceding (un)fair treatment was intentional versus not?

[3] Acknowledging that fairness is a relative term - ranging from low to high level of fairness, throughout the chapter we refer to less fair treatments as "unfair", and more fair treatments as "fair".

Dishonest helping and harming after (un)fair treatment

We consider two possibilities regarding the prevalence of dishonest helping and harming after (un)fair treatment. The first possibility is that dishonest harming after unfair treatment is more prominent than dishonest helping after fair treatment. Supporting this possibility is the idea that "bad is stronger than good". Ample research shows that negative experiences have a higher impact on people's perception, emotions, and behavior than positive experiences (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Brickman, Coates, & Janoff-Bulman, 1978; Rozin & Royzman, 2001). For instance, Klein and Epley (2014) found that people judge prosocial and selfish actions in an asymmetric manner. Specifically, when participants read about someone who engaged in an extremely selfish behavior (e.g., giving much less than a suggested amount to a charity, making a selfish monetary splits in a dictator game) they evaluated the person as less warm and caring, and the behavior as less nice compared to participants who read about someone who engaged in a fair behavior (e.g., giving the suggested amount to a charity, making an equal monetary split in a dictator game). The gap in evaluations between an individual who engaged in fair versus extremely prosocial behavior (e.g., giving twice as much to a charity, giving all the money in a dictator game) was much smaller.

When it comes to reactions to (un)fair treatment, people harm those who were unfair to them more than they benefit those who were fair or even extremely generous to them (Offerman, 2002). Kube, Maréchal, and Puppe (2006) hired students to log information about books into a computer, earning around 15 Euros per hour. When arriving to complete the task, some students were informed that they would be paid 15 Euros per hour, as they expected (neutral treatment). Others, however, were informed that they would earn 10 Euros per hour (unkind treatment) or 20 Euros per hour (kind treatment). Workers in the kind treatment logged 10% more books than workers in the neutral treatment, whereas workers in the unkind treatment logged 27% fewer books than those in the neutral treatment. If people's asymmetric reactions to fair and unfair treatment extend also to situations in which they can affect others by lying, we should expect dishonest harming after unfair treatment to be more prominent than dishonest helping after fair treatment.

The second possibility is that dishonest helping after fair treatment is more prominent than dishonest harming after unfair treatment. Prior work shows that people do not like to actively harm others and will avoid doing so if possible (the do-no-harm principle; Baron, 1995; Van Beest, Van Dijk, De Dreu, & Wilke, 2005). In a series of studies, participants had

to choose between taking an action that will help one group and simultaneously harm another group, or not taking an action at all. Demonstrating the do-no-harm principle, participants were reluctant to choose action over inaction. People preferred to avoid an action that helped group A if it simultaneously harmed group B. This was even the case when the harm to group B was less severe than the harm of not helping group A (Baron, 1995). Moreover, even when two groups are in a conflict, individuals prefer to use their resources to benefit their own group rather than harm the other group (in group love vs. out group hate, Halevy, Weisel, & Bornstein, 2012; Halevy, Kreps, Weisel, & Goldenberg, 2015).

Further supporting the possibility that people react dishonestly more to fairness (versus unfairness), recent work revealed that dishonest helping is seen in a rather positive light. When judging dishonest behavior aimed at helping and harming others, participants evaluated dishonest helping as more acceptable and ethical than dishonest harming (Gino & Pierce, 2010a). Similarly, dishonest helping was perceived as even more ethical than selfish truth-telling (Levine & Schweitzer, 2014). Given that people are averse to harming others, and find dishonest helping rather acceptable, it might be the case that dishonest helping after fair treatment is more prominent than dishonest harming after unfair treatment.

Beyond assessing the relative propensity of dishonest reactions to fair and unfair treatment, we also aim to gain insight into the role of emotions associated with these dishonest reactions. The focal emotions we focus on are gratitude, anger, and disappointment as these emotions are triggered by (un)fair treatment and were found to elicit reactions to it. Prior work has shown that experiencing prosocial gestures like altruism, helping behavior, and fairness increases gratitude, which in turn facilitates helping behavior (Bartlett & DeSteno, 2006; McCullough, Kimeldorf, & Cohen, 2008; Tsang, 2006; 2007). In addition, experiences of unfairness trigger anger (Pillutla & Murnighan, 1996; Seip, Van Dijk, & Rotteveel, 2014) and disappointment (Reuben & van Winden, 2008), which are then associated with subsequent harming behavior. We thus test the extent to which dishonest helping and harming after (un)fair treatment is associated with gratitude, anger, and disappointment.

Overview of studies

We study the prevalence of dishonest helping and harming in three experiments. In all experiments participants first take part in a dictator game. Then, after receiving a less fair (here: unfair) or more fair (here: fair) monetary split from the dictator, recipients engaged in a task in which they could dishonestly inflate or deflate the dictator's pay. In all experiments,

recipients could affect only their counterpart's pay, but not their own pay. Further, in all experiments dictators were not aware of whether or not the recipients engaged in dishonest behavior to affect their pay. Thus, we capture dishonest helping and harming behavior that is removed from any motivation to benefit oneself or motivation to convey an explicit message to the dictator (e.g., teach the dictator a lesson).

In Experiment 3.1 we employed a task that allowed us to assess dishonesty only at the group level. We further assessed a benchmark of dishonest helping/harming when participants did not experience any prior treatment. In Experiment 3.2 we employed a task that allowed us to detect dishonesty at the individual level and test the prevalence of dishonest harming and helping after (un)fair treatment. In a pre-registered Experiment 3.3 we tested whether the source of (un)fairness affects recipients' behavior by adding a control condition in which (un)fairness was determined by a random device, rather than by the dictator. We report all measures, manipulations, and exclusions in the main text and the supplementary online materials (SOM)[4]. The pre-registration for Experiment 3.3[5], as well as all the instructions, manipulations, measures, and data are available on Open Science Framework[6].

Experiment 3.1

Experiment 3.1 tested recipients' dishonest helping and harming after (un)fair treatment. The (un)fair treatment was intentional and created by a dictator in a dictator game. The experiment included several benchmarks for comparison. First, we included a control condition in which participants did not experience any prior treatment at all, but rather could lie to affect an unrelated person's pay. Second, to assess the robustness of any observed findings, we compared participants' behavior when the (un)fair treatment was presented in the form of a gain versus a loss (give-some vs. take-some setting; Krupka & Weber, 2013; Van Dijk & Wilke, 2000). Because losses loom larger than gains (Kahneman & Tversky, 1979), we assess whether being treated unfairly by losing an amount of money evoked stronger dishonest reactions compared with being treated unfairly by not gaining an amount of money.

[4] http://journal.sjdm.org/19/190419/supp.pdf

[5] https://osf.io/zw9ds/?view_only=cd96ba8de56b47e39126f71ac61ebff5

[6] https://osf.io/dyfpu/?view_only=44bf455fcb18496da6e40ee354e61fc9

Method

Participants and procedure. The experiment was conducted on Amazon Mechanical Turk (MTurk), and each participant received a participation fee of 15 cents and an opportunity to earn extra pay. First, we collected dictators' monetary split decisions. Then, we matched a recipient to each dictator, implemented the monetary splits and assessed recipients' behavior. For recipients, the overall design was a 2 (Framing: give-some vs. take-some) × 2 (Amount: unfair vs. fair) + 1 (no prior treatment) between-subjects design. All conditions were run simultaneously, and each participant was randomly assigned to one of the five conditions.

To determine a minimum cell size, we conducted a priori power calculations using G*Power 3.0.10 software with .05 criterion of statistical significance, and 80% power. Since our main goal was to test for differences between participants reactions to unfair and fair treatment, we calculated our sample size focusing on a main effect for fairness. Specifically, we assumed a difference of 15% in recipient reports between the fair and unfair conditions. The calculation indicated that responses from 173 recipients in each cell would be sufficient. Thus, when collecting dictators' decisions, we predetermined that we would stop data collection when we had at least 173 decisions in each of the four (take-some vs. give-some by unfair vs. fair) between-subject cells.

First, dictators were randomly assigned to a give-some versus take-some condition and were asked to split 30 cents (in addition to their participation fee) between themselves and a counterpart (recipient). When making their decisions, dictators were not informed what the second part of the experiment would be. Dictators in the give-some condition were told they had received 30 cents and were then asked to decide between (1) keeping 30 cents for themselves and *giving* 0 cents to their counterpart or (2) keeping 15 cents for themselves and *giving* 15 cents to their counterparts. Dictators in the take-some condition were told their counterpart had received 30 cents and were asked to choose between (1) *taking* 30 cents from the counterpart for themselves and leaving their counterpart with 0 cents or (2) *taking* 15 cents from the counterpart for themselves and leaving their counterpart with 15 cents. In both settings, the decision was identical in monetary terms but different in terms of framing (Krupka & Weber, 2013; Van Dijk & Wilke, 2000). Based on our data collection stopping rule, we stopped collecting dictators' decisions when the smallest cell (take-some, unfair split) reached 176 participants. In total, we collected 1,282 dictators' decisions. Doing so allowed us to follow APA guidelines and avoid deceiving participants, as deception was not necessary.

As a result, our design includes real decisions made by both dictators and recipients. The only trade-off is the unequal cell size.

We then matched each dictator with a recipient. Recipients learned that they were matched with a counterpart (i.e., the dictator), who had decided how to split 30 cents of their own money (give-some) or 30 cents the recipient had initially got (take-some). Recipients learned about the two monetary splits that their counterpart had to choose from (15 cents for each vs. keeping/taking all 30 cents for self) and the decision their counterpart had made. Recipients then engaged in a task in which they could (dishonestly) affect the counterpart's additional pay. Specifically, they were asked to toss a coin and report whether the outcome was "heads" or "tails." (Shalvi, 2012). Reporting "heads" would yield a payoff of an extra 8 cents for their counterpart, whereas reporting "tails" would yield a payoff of an extra 2 cents for their counterpart. Participants were asked to either use a coin they had at home or go to an independent coin-tossing website to which we provided a link. As such, we could not identify participants' actual coin-toss outcome, and could only assess dishonesty at the group level. After the task, recipients assessed the extent to which they thought the amount they received was fair on 7-point scales (1 = not at all, 7 = very much). Further, recipients evaluated the extent to which they were motivated by feelings of gratitude, anger, and disappointment when they completed the coin tossing task. We further collected additional measures for exploratory reasons; see SOM for items and results of these additional measures.

Lastly, in the no prior treatment condition (n = 200), participants engaged in the coin-tossing task affecting the payoff of a counterpart without the first dictator stage. They further evaluated their motivations on the same scales as the recipients in the other conditions. In total, we collected data from 2,764 participants (1,282 dictators, 1,282 recipients, and 200 participants in the no prior treatment condition; M_{Age} = 36.63, SD_{Age} = 12.13; 60.7% female; gender had no effect on behavior, see SOM). After collecting all of the data, all participants were paid according to their own and their counterpart's actual decisions.

Results

In the give-some condition, 347 dictators (62.97%) choose the fair split (giving 15 out of 30 cents), whereas 204 dictators (37.03%) chose the unfair split (giving 0 out of 30 cents). In the take-some condition, 555 dictators (75.92%) choose the fair split (taking 15 out of 30 cents), whereas 176 dictators (24.08%) chose the unfair split (taking 30 out of 30 cents). Replicating prior findings (Krupka & Weber, 2013; Van Dijk & Wilke, 2000), dictators

choose more fair allocations in a take-some (75.92%) than give-some (62.97%) setting, $\chi^2(1) = 25.25$, $p < .001$, Cramer's $V = 0.140$.

Fairness. A two-way ANOVA with the Amount (unfair [0 cents] vs. fair [15 cents]) and Framing (give-some vs. take-some), predicting the extent to which recipients evaluated the amount they received as fair, revealed a main effect of the amount. Participants who received 15 cents evaluated the amount they received as fairer ($M = 6.02$, $SD = 1.43$) than those who received 0 cents ($M = 2.63$, $SD = 1.95$), $F(1, 1278) = 1{,}237.40$, $p < .001$, $\eta^2 = .492$. This was not qualified by an Amount × Framing interaction, $p = .936$.

Recipients' behavior. A chi-squared analysis revealed that, overall, the proportion of reported "heads" was higher among participants who received a fair monetary split (62.08%) than among participants who received an unfair monetary split (52.63%), $\chi^2(1) = 9.89$, $p = .002$, Cramer's $V = .088$. A binomial test showed that the proportion of "heads" after being treated fairly was higher than the 50% expected from honest reports, $p < .001$. By contrast, the proportion of "heads" after unfair treatment did not differ from 50%, $p = .330$.

A log-linear analysis revealed that the Amount × Framing interaction predicting the reported coin-toss outcome ("heads" vs. "tails") was not significant, $p = .715$, indicating that in our setting, the give-some vs. take-some framing did not affect recipients' behavior. Indeed, out of those who received a fair amount, 61.95% reported "heads" in a give-some framing and 62.16% in a take-some framing. Of those who received an unfair amount, 51.47% reported "heads" in a give-some framing, and 53.97% in a take-some framing. Because the framing did not affect participants' behavior, in the remaining analyses we collapsed the two framing conditions.

No prior treatment. Participants in the no prior treatment condition reported "heads" in 63.00% of the cases, which was significantly higher than 50%, $p < .001$. Post-hoc analysis showed a significant difference between participants in the no prior treatment condition (63.00% "heads") and those who received an unfair amount from their counterparts (52.63% "heads"), $\chi^2(1) = 5.72$, $p = .017$, Cramer's $V = .099$. We found no difference between participants in the no prior treatment condition and those who received a fair amount from their counterparts (62.08% "heads"), $p = .809$.

Emotions. We assessed whether different reports ("heads" vs. "tails") when participants were treated unfairly, fairly, or not treated at all were associated with different levels of emotions. Because emotions were always measured after participants' coin-toss

reports, we test for only association between dishonest reactions and emotions, and refrain from making causal claims. We thus ran a series of ANOVAs with 2 (Report: "heads" vs. "tails") by 3 (Condition: unfair vs. fair vs. no prior treatment) predicting gratitude, anger, and disappointment. Results revealed a main effect for Condition, for gratitude, $F(2, 1476) = 115.11$, $p < .001$, $\eta^2 = .135$; anger, $F(2, 1476) = 70.22$, $p < .001$, $\eta^2 = .087$; and disappointment, $F(2, 1476) = 134.71$, $p < .001$, $\eta^2 = .154$. Overall, participants reported higher levels of anger and disappointment and lower gratitude in the unfair condition than in the fair and no prior treatment conditions, ps < .001. There was no difference in the levels of anger, disappointment, and gratitude between the fair and no prior treatment conditions, ps > .262.

Additionally, there was a significant Report × Condition interaction for gratitude only, $F(2, 1476) = 5.09$, $p = .006$, $\eta^2 = .007$. Simple effects revealed that, among participants who were treated unfairly, those who reported an outcome that does not benefit the dictator ("tails") reported lower levels of gratitude than those who reported an outcome that benefits the dictator ("heads"), $p < .001$. Participants who were treated fairly or not treated at all reported similar levels of gratitude regardless of whether they reported an outcome that benefits their counterpart or not; see Table 3.1 and SOM for detailed analyses.

Table 3.1. *Self-reported emotions in Experiment 3.1.*

	n	Gratitude	Anger	Disappointment
Unfair amount (0 cents)				
Reporting "heads"	200	3.92 (1.42)	2.48 (1.91)	3.09 (2.14)
Reporting "tails"	180	3.28 (1.52)	2.51 (1.93)	3.15 (2.21)
Difference		***		
Fair amount (15 cents)				
Reporting "heads"	560	4.93 (1.28)	1.56 (1.21)	1.63 (1.31)
Reporting "tails"	342	4.79 (1.35)	1.50 (1.11)	1.66 (1.27)
Difference				
No prior experience				
Reporting "heads"	126	4.73 (1.31)	1.46 (1.10)	1.46 (1.06)
Reporting "tails"	74	4.70 (1.38)	1.36 (0.91)	1.40 (0.92)
Difference				

Note. Means (SDs) of the level of gratitude, anger, and disappointment 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: $^{***}p < .001$. When adjusting significance level for all the measures we collected (7 in total, see SOM), the new significance level is 0.05/7 = 0.007. $p < .007$ will be considered significant, thus all comparisons marketed as *** remain significant.

Discussion

Results of Experiment 3.1 reveal that after being treated fairly or not being treated at all, people, on average, over-report coin-toss outcomes to benefit their counterparts. However, when people experience unfair treatment from their counterparts, they, on average, neither over- nor under-report coin toss outcomes. Overall, these results are consistent with Klein and Epley (2014) and seem to suggest that when people can engage in dishonest behavior, they react more to unfairness, rather than to fairness. Compared to a baseline of no prior treatment, people do not adjust their behavior when they were treated in a fair manner, as evident by the similar proportion of "heads" in the no prior treatment and fair treatment condition. However, compared to no prior treatment, people do adjust their behavior after being treated unfairly. They seem to engage in less dishonest helping and on average report a proportion of "heads" that is not different than the proportion of "heads" expected from an honest report.

Interestingly, we did not detect differences between participants' fairness evaluations, as well as their dishonest reactions to fair and unfair treatments in a give-some versus take-some framing. There are two potential reasons for this lack of difference (as opposed to prior work in which differences were observed; Krupka & Weber, 2013; Van Dijk & Wilke, 2000). First, it might be the case that our sample size was not sufficient to detect the effect. A sensitivity analysis for our sample size (n = 1,282), with 80% power to detect an effect and significance level of .05 suggests that our sample was sufficient to detect a small size effect (Cramer's V = .078). Our a-priori power calculation, however, focused on detecting differences between the fair and unfair conditions and not on detecting an Amount × Framing interaction. Further, assessing the social appropriateness of dictator splits, Krupka and Weber (2013) find a rather small difference in people's evaluations when the same monetary split is framed as a give-some versus take-some[7]. Consequently, it might be that our sample size was not sufficient to detect such a subtle effect.

Second, in our task, participants in the take-some conditions were not physically endowed with the payoff. Although participants read that they were endowed with the payoff, the payoff was tangibly given to them only at the end of the study. It might be that paying participants upfront and then actually taking some payment away, or providing a visual representation of money being taken away would have elicited stronger reactions in the take-some conditions. Since the take-some versus give-some manipulation is not the focal point of the current work, in Experiment 3.2 and 3.3 we focused on give-some framing only.

In the coin-tossing task, recipients could report honestly (i.e., report the coin-toss outcome they actually observed), lie to help their counterpart (i.e., see "tails" but report "heads"), or lie to harm their counterpart (i.e., see "heads" but report "tails"). Since we do not know what recipients' actual coin toss outcome was, we cannot determine for every individual participant whether he or she was honest, lied to help, or lied to harm their counterpart. As such, there are two ways in which the ~50% of "heads" in the unfair treatment condition can be interpreted. One possibility is that most of the participants who were treated unfairly honestly reported the coin toss outcome they saw, and only few (if any) lied. Such behavior will suggest that after being treated unfairly, people cease lying to help others, but do not start lying to harm others. The second possibility is that, after being treated unfairly,

[7] Krupka & Weber (2013) find a gap of 0.10 in evaluations of social appropriateness for a dictator split of 100%-0% when it is in a give-some vs. take-some framing. They find a gap of 0.06 in evaluations of social appropriateness for a dictator split of 50%-50% when it is in a give-some vs. take-some framing.

some participants lied to harm their counterpart, but others lied to help the counterpart. If the proportion of dishonest helpers and harmers is similar, the overall proportion of participants reporting "heads" will be close to ~50%. In order to be able to distinguish between participants who dishonestly helped and harmed their counterparts, as well as accurately detect the emotions associated with dishonest harming and helping after (un)fair treatment, in Experiment 3.2 and 3.3 we employed a task that allows assessing dishonesty at the individual level.

Experiment 3.2

As in Experiment 3.1, in Experiment 3.2 we assessed dishonest helping and harming following experiencing an intentional (un)fair treatment in a dictator game. Experiment 3.2 employed a task that allows classifying individuals into honest, dishonest helpers, and dishonest harmers. Further, since Experiment 3.1 established that dishonest helping is similar after fair treatment and no treatment, in Experiment 3.2 we focus on comparing between fair and unfair treatments only.

Method

Participants and procedure. Participants arrived at the lab in an Israeli university in groups of 6 to 24, to complete an experiment in exchange for course credit and an opportunity to earn extra money. Our predetermined data collection stopping rule was to collect as much data as possible during the semester and stop when the semester was over. In the time allocated for running the study, we were able to collect data from 160 participants (M_{Age} = 23.66, SD_{Age} = 1.52; 79.37% females, gender had no effect on behavior, see SOM), leading to a total of 80 × 96 = 7680 observations. A sensitivity power analysis with a .05 criterion of statistical significance and a power of 80% showed we had power to detect a medium effect of f = .31 with a sample size of 80 receivers.

Participants were randomly assigned to the role of dictator (n = 80) or recipient (n = 80), and were randomly paired with a counterpart whose identity remained anonymous throughout and after the experiment. Dictators received 20 ILS (1 ILS = ~€0.25) in 2 ILS coins, and were asked to split the money between themselves and the recipient. They were asked to place the amount they chose to keep for themselves in an envelope labeled "To me" and the remaining amount in an envelope labeled "To the other person." When making their decisions, dictators were not informed what the second part of the experiment would be. The

experimenter then transferred the envelope labeled "To the other person" to a respective recipient who was seated in another room.

In turn, recipients received their envelope, opened it, and learned how much money their counterpart had decided to give them. Recipients were aware of the possible monetary splits the dictators could choose from. Recipients further learned they would engage in a task, and that their performance in the task would affect their counterparts' additional pay. In the task (the ambiguous die paradigm; taken from Pittarello et al., 2015[8]), recipients were presented with a fixation cross that appeared on a computer screen for 1000ms. After the fixation cross disappeared, six die-roll outcomes appeared for 2000ms. Participants were asked to report the die outcome that appeared closest to the preceding fixation cross (i.e., target); see Figure 2.2. Participants engaged in this task for 196 trials, each time reporting the die outcome that appeared closest to the fixation cross. Participants were informed that, after completing the task, one trial would be randomly chosen and the outcome reported on that trial would determine their counterpart's payoff, with higher outcomes corresponding to higher payoffs (i.e., reporting 1 means the dictator earns 2 ILS, 2 = 4 ILS, 3 = 6 ILS, 4 = 8 ILS, 5 = 10 ILS, and 6 = 12 ILS).

Out of 196 trials, 96 were experimental trials. In those trials, the outcome actually closest to the fixation cross (target) was '3'. Across trials, we varied the value second-closest to the fixation cross (i.e., value next to the target) to be higher (i.e., 4 or 5; helping the dictator) or lower (i.e., 1 or 2; harming the dictator) than the target outcome. In the experimental trials, '3' was always the target outcome in order to allow keeping the absolute gap between the target outcome and the value next to the target constant (i.e., a gap of 1 between 2 and 3, or 4 and 3; a gap of 2 between 1 and 3, or 5 and 3). We further varied the location of the target die (second vs. third vs. fourth vs. fifth die from the left) and the location of the fixation cross (20 vs. 40 vs. 60 pixels away from the target die). The fixation cross was always objectively closer to the target than to the value next to the target. To diversify the values that participants saw as the target, in the additional 100 filler trials the target outcome was 1, 2, 4, 5, or 6. We report additional analysis on the effect of gap, target location, fixation-cross location, and trial number on recipients' behavior in the SOM.

The task allowed assessing dishonest helping and harming at an individual level. If participants wish to be honest, they should report the correct value, 3, in the majority of the

[8] Available at https://osf.io/8hbti/

trials. Participants are indeed able to do that rather well when incentivized to be accurate (see Pittarello et al., 2015). Further, when participants are motivated to be honest, but misreport, they are likely (1) to report the value second-closest to the target (i.e., make a mistake) and (2) to not systematically misreport values that are mostly higher or lower than 3 (see Pittarello et al., 2015). That is, the proportion of misreports that help the counterpart (i.e., reports > 3) should be similar to the proportion of misreports that harm their counterpart (i.e., reports < 3). However, if participants wish to dishonestly harm (or help) their counterpart, they may (1) report any value even if it did not appear close to the target (or even did not appear on the screen at all), and (2) make systematic reports that mostly harm or help their counterparts. By systematically harming (helping) their counterparts, recipients can ensure their counterpart will be paid less (more) than she is supposed to if participants complete their task accurately.

After reading the instructions, recipients completed three practice trials, followed by a comprehension question. Specifically, they completed the following sentence "My performance in the following task will determine" by choosing between (1) "how much I will earn", and (2) "how much my counterpart will earn." Recipients then completed a first block of 98 trials followed by a reminder of the payoff structure, and a second block containing the remaining 98 trials. Upon completion of the task, recipients assessed the extent to which they though the amount they received was fair on a 7-point scale (1 = not at all, 7 = very much). As in Experiment 3.1, recipients evaluated the extent to which their behavior in the task was driven by gratitude, anger, and disappointment. As in Experiment 3.1, additional items for exploratory reasons were also assessed (see SOM for details and analyses.)

Results

Two participants answered the comprehension question incorrectly and four participants did not answer it at all. We excluded those six participants from all analyses, leaving a sample of 74 participants (7,104 observations, nested within participants). Including these participants did not change the obtained results. Recipients received amounts that ranged between 0 and 20 ILS. Two recipients (2.7%) received 20 ILS, and the rest got an amount between 0 and 10 ILS. The average amount received was 7.56 ILS (*SD* = 3.66), and 87.83% of the amounts were within ±1 *SD* from the mean.

Fairness. The higher the amount participants got, the fairer they evaluated the amount to be, $r = .57$, $p < .001$.

Recipients' behavior. Table 3.2 presents the proportion of each outcome recipients reported as a function of the value that appeared next to the target. As can be seen, participants reported the correct target value, 3, in 62.24% of the trials, and misreported the target in the remaining 37.76% of the trials. Participants reported the value next to the target in 63.38% of the misreported trials, whereas in 36.62% of the misreported trials they reported a different value. Due to the considerable proportion of trials in which participants misreported the target, but reported a value that cannot be perceived as a "mistake" (that is, a value that is not next to the target), in the remaining analyses we analyze participants' reports based on all their incorrect responses (i.e., all values that differ from the target)[9].

Table 3.2. *Reported outcomes in Experiment 3.2.*

Value next to the target	Participants' reported value (%)						
	1	2	3	4	5	6	Other
1	**13.86%**	3.24%	*73.98%*	0.91%	2.22%	5.80%	0.00%
2	1.77%	**17.67%**	*71.04%*	1.20%	1.77%	6.56%	0.00%
4	2.33%	4.38%	*53.75%*	**31.31%**	1.59%	6.59%	0.06%
5	2.95%	3.58%	*52.67%*	1.42%	**33.77%**	5.62%	0.00%
Total	5.18%	7.14%	*62.24%*	8.63%	9.75%	6.08%	0.01%

Note. The proportion of the reported value per the value next to the target. The proportion of the correct value, 3, is in italics. The proportion of reports of the value next to the target is in bold. "Other" represents reporting other values (e.g., typos).

We classified every report that was higher than 3 (the correct value) as helping the counterpart, and every report that was lower than 3 as harming the counterpart. Notably, the likelihood of reporting an outcome that helps versus harms the counterpart was not associated with the trial number, $p = 0.79$, suggesting that the misreports were not driven by boredom or fatigue due to the length of the task.

Further, for each participant we counted the number of trials in which they reported (1) the correct value, 3, (2) a value that is higher than 3 (helping the counterpart), and (3) a

[9] We thank the reviews for suggesting to include all reports in the analyses, instead of focusing only on the reports of the value next to the target. Analyzing the data taking into account only reports of the value next to the target (which can be perceived as other-helping and other-harming "mistakes") did not change the results.

value that is lower than 3 (harming the counterpart). We then calculated for each participant the proportion of misreports that help the dictator out of the total number of misreports. This proportion ranges from 0% for a participant who made only other-harming misreports to 100% for a participant who made only other-helping misreports. A proportion of 50% indicated that a participant made the same number of helping and harming misreports, thus not systematically helped or harmed her counterpart. A linear regression with the amount participants received, predicting the proportion of helping misreports (out of all misreports), revealed that the higher the amount participants received, the higher the proportion of helping misreports (and thus lower the proportion of harming misreports), $b = 0.026$, $t(72) = 2.72$, $p = .008$.

We then classified participants into behavioral types, according to whether they made systematic, intentional misreports that harmed or helped their counterpart. To do so we compared the proportion of helping misreports (out of all misreports) to a binomial distribution of 50% misreports that help and 50% misreports that harm (the expected pattern if participants' misreports are unintentional). Participants whose misreports did not differ from the binomial distribution were classified as "inconsistent". Participants whose proportion of helping misreports was significantly higher than 50% were classified as "dishonest helpers", and participants whose proportion of helping misreports was significantly lower than 50% were classified as "dishonest harmers". Overall, 25 recipients (33.78%) were classified as inconsistent. A total of 39 (52.70%) were classified as dishonest helpers, and the remaining 10 (13.51%) were classified as dishonest harmers. A chi-squared analysis revealed that the behavioral types (dishonest helpers vs. dishonest harmers vs. inconsistent) differed across the amounts participants received (0 vs. 2 vs. 4 vs. 6 vs. 8 vs. 10 vs. 20), $\chi^2(12) = 34.41$, $p < .001$, Cramer's $V = .482$.

In addition to assessing participants' dishonest helping and harming behavior as a function of the continuous amount they received, we wanted to assess dishonest reactions to (un)fair treatment. To do so, we had to decide on a cutoff point from which an amount of money will be classified as fair or unfair. One potential cutoff is below versus above 50% of the initial endowment, such that all amounts below 10 ILS will be considered unfair, whereas 10 ILS and above it will be considered as fair. Such classification is reasonable because splitting the endowment equally is the second most common split dictators choose to implement (the most common is not giving any amount; Engel, 2011). Figure 3.1 presents the fraction of dishonest helpers, dishonest harmers, and inconsistent participants as a function of

whether participants received a fair amount (10 ILS or more) or unfair amount (less than 10 ILS). As can be seen, out of all the participants who received a fair amount (n = 33), 72.72% were dishonest helpers and the remaining 27.27% were inconsistent. Out of all the participants who received an unfair amount (n = 41), 24.39% were dishonest harmers, and 39.02% were inconsistent. Interestingly, even after an unfair amount, 26.58% of the participants were dishonest helpers.

A chi-squared analysis employing the 50% of endowment cutoff (unfair: 0-8 ILS, fair: 10-20 ILS) revealed that the proportion of dishonest helpers among participants who received a fair amount (72.72%) was higher than the proportion of dishonest harmers among participants who received an unfair amount (24.39%), $\chi^2(1) = 15.31$, $p < .001$, Cramer's V = .482. However, the meaningful proportion of dishonest helpers even among participants who received an unfair amount (36.58%) seems to suggest that people are likely to engage in dishonest helping even after being treated unfairly.

When dictators were free to give any amount they wished from their initial endowment, there is no inherent cutoff point from which an amount should be considered as fair or unfair, and the 50%-50% cutoff was determined based on previous research (Engel, 2011). We further assessed whether our results robust to alternative cutoff points for fairness. In the SOM we report additional analyses employing the (1) median and mean split of the amount participants received as a cutoff for fairness, (2) a median split of participants' subjective evaluation of fairness, and (3) a cutoff of 30% of the initial endowment (a commonly accepted monetary split in an ultimatum game; Güth, Schmittberger & Schwarze, 1982). Analyses reveal that our results are robust to all alternative cutoff points of fairness, see SOM. Note that the existence of dishonest helpers among those who received an unfair amount was not limited to the categorization of unfairness as less than 50% of the endowment. Figure S1 in the SOM presents the behavioral types as a function of the amount recipients received as a continuous measure, showing that for every small amount participants received (0 ILS, 2 ILS, 4 ILS, and 6 ILS) some of the recipients were classified as dishonest helpers, see SOM for details.

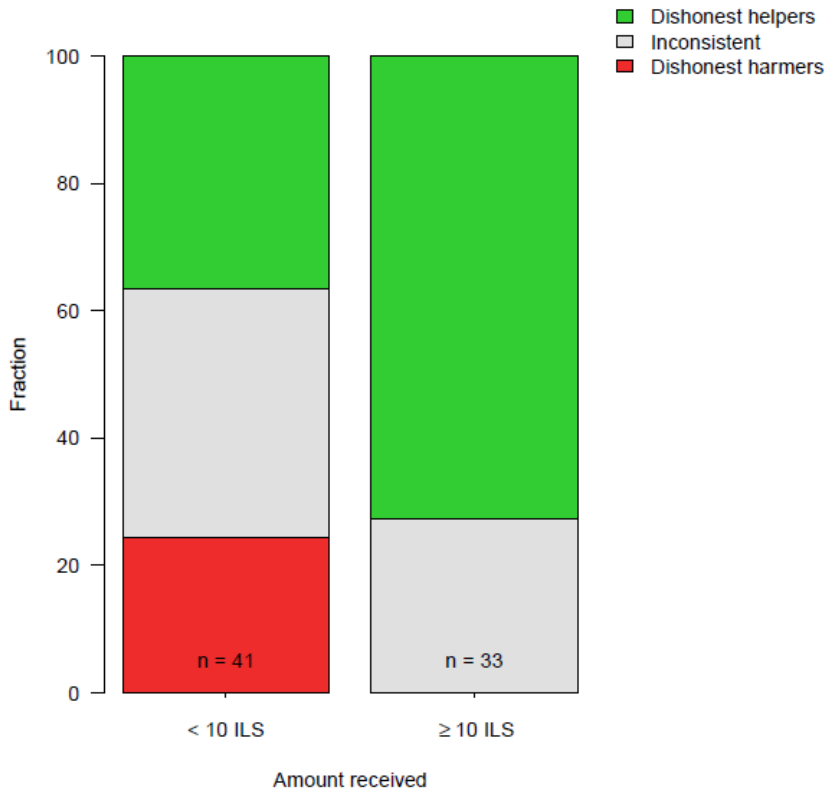

Figure 3.1. The fraction of behavioral types (dishonest helpers, dishonest harmers, and inconsistent), as a function of the amount participants received (0-8 ILS; 10-20 ILS). The *N*s of each group appear on the bar.

Emotions. We further assessed whether the different behavioral reactions to similar amounts of money were associated with different levels of gratitude, anger, and disappointment. Because emotions were measured only after participants' behavior, we assess only the association between dishonest reactions to (un)fair treatment and emotions and refrain from making causal claims. We particularly tested whether, after receiving unfair amounts, participants who engaged in dishonest harming reported different levels of anger, disappointment, and gratitude than those who did not dishonestly harmed their counterparts. Similarly, we tested whether, following a fair amount, those who engaged in dishonest helping reported different levels of these emotions than those who did not dishonestly helped their counterparts. A series of ANOVA analyses with 2 (Behavior: dishonest helping after being treated fairly/dishonest harming after being treated unfairly vs. a different reaction to being treated (un)fairly) by 2 (Amount: unfair [0-8 ILS] vs. fair [10-20 ILS]) predicting gratitude, anger, and disappointment revealed a main effect for Amount on gratitude, anger, and disappointment. Participants who received fair amounts, reported overall more gratitude (M = 5.75, SD = 1.27) than those who received unfair amounts (M = 3.97, SD = 1.60), F(1,

70) = 59.50, $p < .001$, $\eta^2 = .459$. Further, those who received fair amounts were less angry ($M = 1.09$, $SD = 0.38$) than those who received unfair amounts ($M = 2.15$, $SD = 1.75$), $F(1, 70) = 74.85$, $p < .001$, $\eta^2 = .517$. Similarly, those who received fair amounts were less disappointed ($M = 1.15$, $SD = 0.44$) than those who received unfair amounts ($M = 2.34$, $SD = 1.66$), $F(1, 70) = 65.80$, $p < .001$, $\eta^2 = .485$.

Importantly, the Behavior × Amount interactions was significant for gratitude, anger, and disappointment, Fs(1, 70) > 9.86, ps < .002, η^2s > .124. Assessing simple effects revealed that participants who received fair amounts reported similar levels of gratitude, anger, and disappointment regardless of whether they engaged in dishonest helping or not. In contrast, those who received unfair amounts reported different levels of gratitude, anger, and disappointment depending on whether they dishonestly harmed their counterpart or not. Those who engaged in dishonest harming reported that they were angrier, more disappointed, and less grateful than those who did not engage in dishonest harming; see Table 3.3. We report all detailed analyses in the supplementary materials (SOM). Employing alternative cutoff points for fairness yielded the same results.

Table 3.3. *Self-reported emotions in Experiment 3.2.*

	n	Gratitude	Anger	Disappointment
Unfair amounts (0-8 ILS)				
Dishonest harming	10	2.06 (0.76)	4.60 (1.57)	4.50 (1.43)
No dishonest harming	31	4.58 (1.28)	1.35 (0.83)	1.65 (1.01)
Difference		***	***	***
Fair amounts (10-20 ILS)				
Dishonest helping	24	5.62 (1.31)	1.13 (0.44)	1.17 (0.48)
No dishonest helping	9	6.08 (1.18)	1.00 (0.00)	1.11 (0.33)
Difference				

Note. Means (SDs) of the level of gratitude, anger, and disappointment, 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: *** $p < .001$. When adjusting significance level for all the measures we collected (5 in total, see SOM), the new significance level is 0.05/5 = 0.01. $p < .01$ will be considered significant, thus all comparisons marked as *** remain significant.

Discussion

In Experiment 3.2 we employed the ambiguous die paradigm, which allowed us to assess dishonest helping and harming at an individual level. Our results reveal that dishonest helping behavior is rather common. Both when people were treated fairly, as well as when they were treated unfairly, a non-negligible proportion of individuals dishonestly helped their (un)fair counterpart. Only when treated unfairly, some individuals engaged in dishonest harming. The rather high prevalence of dishonest helping, regardless of a preceding fair or unfair treatment is in line with the results obtained in Experiment 3.1, in which participants dishonestly helped those they had never interacted with before. Taken together, results of both experiments seem to suggest that dishonest helping is a rather robust and common behavior. Only after unfair treatment, some people stop dishonestly help and even start to dishonestly harm their unfair counterpart. We interpret these findings as suggesting that when people can react in dishonest means, they react more to a rather unfair treatment, compared to a rather fair treatment.

Assessing participants' emotions correspondingly points to a higher sensitivity to unfairness, compared to fairness. In particular, while among participants who were treated fairly, there was no association between their dishonest reactions and emotions, such differences were apparent among participants who were treated unfairly. In particular, after unfair treatment, dishonest harming was associated with higher levels of anger and disappointment, and lower levels of gratitude than a different reaction to the unfair treatment.

Experiment 3.3

One question that remains open is whether participants dishonest reaction to (un)fair treatment is driven by, or independent of, their motivation to reciprocate their (un)fair counterpart? In order to test the role of the motivation to reciprocate, in a pre-registered Experiment 3.3 we added a control condition in which the monetary split between the recipients and their counterpart was determined randomly. Thus, in Experiment 3.3, recipients received a fair or unfair amount that was determined either by a dictator or randomly. After receiving the amount of money, recipients engaged in the same task as in Experiment 3.2, affecting only their counterparts' pay.

If indeed the motivation to reciprocate drives the behavioral pattern in Experiment 3.2, we should find that (1) after unfair treatment, fewer participants engage in dishonest harming when the allocation was determined randomly versus by the dictator and (2) the level of

gratitude, anger, and disappointment should vary when the amount is determined by a dictator versus randomly. We would not necessarily expect less dishonest helping when the money is determined by a dictator versus randomly, because results of Experiment 3.1 revealed that people are equally likely to dishonestly help others in the absence of prior treatment. Since prior work has found that reactions to (un)fair gestures are stronger when (un)fairness was intentional compared to not (Falk, Fehr, & Fischbacher, 2008; Falk & Fischbacher, 2006, Offerman, 2002), we predicted the pattern mentioned above in our pre-registration.

Alternatively, when people influence others by lying, they may react to the mere feeling of being treated (un)fairly and not to whether the (un)fairness was intentional. If true, we should find (1) a similar level of dishonest harming after being treated unfairly, regardless of whether the unfairness was determined randomly or by a dictator, and (2) a similar pattern of gratitude, anger, and disappointment across both conditions.

Method

Participants and procedure. Based on the results of Experiment 3.2, we conducted an a priori power calculation using G*Power 3.0.10 software to determine the minimum cell sizes for Experiment 3.3 (see pre-registration for details). We used a .05 criterion of statistical significance and 80% power to detect an effect. The calculation indicated that a total sample of 83 recipients in each condition (random vs. dictator) would be sufficient. To stay on the conservative side, we pre-registered that we would collect a total of 100 participants for each condition.

In the first stage, we collected 100 monetary split decisions from dictators on MTurk. All dictators received a fixed payment of 10 cents and were asked to split 20 tokens between themselves and their counterpart. Dictators could choose between (a) keeping 18 tokens for themselves and giving 2 tokens to their counterpart and (b) keeping 10 tokens for themselves and giving 10 tokens to their counterpart. Dictators were aware that each token was worth $0.28 (= 1 ILS) and that their counterpart was someone who would participate in an experiment at our university. When making their decisions, dictators were not informed what the second part of the experiment would be.

After collecting responses from 100 dictators (M_{Age} = 36.61, SD_{Age} = 13.34; 49.0% female), we collected recipients' responses in a computer lab in Israel. A total of 200 participants (M_{age} = 25.74, SD_{age} = 7.10; 47.0% female, gender had no effect on behavior, see SOM) in the lab received 10 ILS for participation, and could earn additional pay based on the

instructions. They were randomly assigned to one of four conditions in a 2 (Amount: unfair [2 ILS] vs. fair [10 ILS]) by 2 (Allocation: dictator vs. random) between-subjects design. Upon arriving at the lab, each participant sat in front of a computer screen and received an envelope with either 2 ILS or 10 ILS in it. First, participants learned that they were paired with a counterpart who participated in the experiment online and was not in the room with them.

Then, participants in the dictator condition learned that, a few days prior to their arrival at the lab, their counterpart decided how to split 20 ILS between them. Recipients learned about the two monetary split alternatives their counterpart had had (10 ILS for each vs. taking 18 ILS for themselves and giving 2 ILS for the recipient). Then recipients were instructed to open the envelope in front of them to learn how much money their counterpart had decided to give them. After learning about the amount, recipients engaged in the ambiguous die paradigm affecting their counterpart's (i.e., the dictator's) pay.

Participants in the random allocation condition learned that a random device determined the monetary split between themselves and their counterpart. As in the dictator condition, recipients in the random allocation condition knew what the two monetary split options were (10 ILS to each vs. 18 ILS for their counterpart and 2 ILS for them). Then participants were instructed to open the envelope in front of them to learn how much money was randomly allocated to them. To keep both settings identical, the distribution of monetary splits in the random condition was identical to the distribution of splits made by the dictators in the dictator condition. Recipients in the random condition were not aware what the exact distribution was. After learning about the amount that was randomly allocated to them, recipients engaged in the ambiguous die paradigm affecting their counterpart's pay. As in Experiment 3.2, we report additional analysis on the effect of different task characteristics (i.e., gap between target and value near the target, target location, fixation-cross location, and trial number) on recipients' behavior in the SOM.

As in Experiment 3.2, recipients were informed that, after completing the task, one trial would be randomly selected and that the outcome reported on that trial would determine their counterpart's payoff, with higher outcomes corresponding to higher payoffs (i.e., reporting 1 means the counterpart earns 1 ILS, 2 = 2 ILS, 3 = 3 ILS, 4 = 4 ILS, 5 = 5 ILS, and 6 = 6 ILS). Lastly, as in Experiment 3.2, after the task participants evaluated the fairness of the amount they received and the same set of scales as in Experiment 3.2 (see SOM for details). After we collected the data, all participants were paid according to their and their counterpart's decisions.

Results

Fairness. A 2 (Amount: unfair [2 ILS] vs. fair [10 ILS]) by 2 (Allocation: dictator vs. random) ANOVA predicting the extent to which participants evaluated the amount as fair revealed a main effect for Amount, $F(1,189) = 367.85$, $p < .001$, $\eta^2 = .665$. Participants who received 10 ILS evaluated the amount as fairer ($M = 6.12$, $SD = 1.59$) than those who received 2 ILS ($M = 1.83$, $SD = 1.52$). The Amount × Allocation interaction was not significant, $p = .213$, indicating that participants evaluated the amount as fair (or not) regardless of how the (un)fairness was determined.

To verify that we had enough power to detect and meaningfully interpret the null interaction, we conducted a sensitivity analysis. The sensitivity analysis for the obtained sample size ($n = 200$), with 80% power to detect an effect and significance level of .05 suggests that our sample was sufficient to detect a medium to small size effect ($f = 0.19$). As such, we interpret the lack of Amount × Allocation interaction on fairness evaluation as suggesting that indeed, participants were not sensitive to whether the (un)fair treatment was determined intentionally, by a dictator, or randomly.

Recipients' behavior. Table 3.4 presents the proportion of each outcome recipients reported per the value that appeared next to the target. As can be seen, participants reported the correct target, 3, in 59.90% of the trials and misreported the target in 40.10% of the trials. Participants reported the value next to the target in 68.58% of the misreported trials, whereas in 31.42% of the misreported trials they reported a different value.

Table 3.4. *Reported outcomes in Experiment 3.2.*

Value next to the target	Participants' reported value (%)						
	1	2	3	4	5	6	Other
1	**17.71%**	0.48%	*70.79%*	0.46%	1.42%	9.07%	0.07%
2	1.69%	**17.76%**	*69.31%*	0.67%	0.96%	9.54%	0.07%
4	1.72%	0.56%	*50.54%*	**36.86%**	0.79%	9.36%	0.17%
5	1.72%	0.46%	*50.09%*	0.44%	**38.21%**	9.02%	0.06%
Total	5.69%	4.79%	*59.90%*	9.55%	10.29%	9.20%	0.06%

Note. The proportion of the reported value per the value next to the target. The proportion of the correct value, 3, is in italics. The proportion of reports of the value next to the target is in bold. "Other" represents reporting other values (e.g., typos).

As in Experiment 3.2, to capture all of participants reports we classified trials in which participants reported a value higher than 3 as helping the counterpart, and trials in which participants reported a value lower than 3 as harming the counterpart[10]. As in Experiment 3.2, the likelihood of reporting an outcome that helps versus harms the counterpart was not associated with the trial number, $p = 0.48$, suggesting that participants' misreports capture intentional dishonesty and not misreports made due to boredom or fatigue.

We then calculated for each participant the proportion of misreports that helped the counterpart out of the total number of misreports. As in Experiment 3.2, the proportion ranges from 0% (a participant who made only other-harming misreports) to 100% (a participant who made only other-helping misreports). A 2 (Amount: unfair [2 ILS] vs. fair [10 ILS]) by 2 (Allocation: dictator vs. random) ANOVA predicting the proportion of misreports that helped revealed a main effect for Amount, $F(1, 195) = 4.06$, $p = .045$, $\eta^2 = .020$. Participants who received 2 ILS had a lower proportion of misreports that helped (and thus a higher proportion of misreports that harmed) out of all misreports ($M = 70.94\%$; $SD = 29.50\%$) than participants who received 10 ILS ($M = 78.72\%$; $SD = 21.68\%$). The Amount × Allocation interaction was not significant, $p = .596$. Bayesian analyses comparing a model with only Amount as the predictor for the proportion of helping misreports with a model that includes the Amount, the Allocation, and an Amount ×Allocation interaction as a predictors revealed a Bayes factor of $BF_{10} = 0.075$, suggesting strong evidence in favor of a model with only Amount as a predictor. Specifically, our data is 13.33 times more likely to occur when Amount is the only predictor for the proportion of helping misreports than when Amount, Allocation, and an Amount x Allocation interaction predict the proportion of helping misreports. Thus, it seems that the source of the (un)fairness, whether determined by a dictator or a random device, did not affect participants' reports.

As in Experiment 3.2, we then classified participants into dishonest helpers, dishonest harmers, and inconsistent by comparing their proportion of misreports that helped the counterpart (out of all misreports) to a binomial distribution of 50% misreports that help, and 50% of misreports that harm, Figure 3.2. Overall, 81 recipients (40.5%) were classified as inconsistent. A total of 107 (53.5%) were classified as dishonest helpers, and the remaining 12 (6%) were classified as dishonest harmers. A chi-squared analysis revealed that the frequency

[10] We thank the reviews for suggesting to include all reports in the analyses, instead of focusing only on the reports of the value next to the target. Analyzing the data taking into account only reports of the value next to the target (which can be perceived as other-helping and other-harming "mistakes") did not change the results.

of types (dishonest helpers vs. dishonest harmers vs. inconsistent) differed among those who received a fair and unfair amount, $\chi^2(2) = 9.44$, $p = .009$, Cramer's $V = .217$. Specifically, among those who received a fair amount (i.e., 10 ILS out of 20), 60.49% were dishonest helpers, 39.50% were inconsistent, and no participants was classified as a dishonest harmers. Among those who received an unfair amount (2 ILS out of 20), 48.73% were dishonest helpers, 10.08% were dishonest harmers, and 41.17% were inconsistent.

A log-linear analysis revealed that the Amount ×Allocation interaction predicting the frequency of types (systematic helpers vs. systematic harmers vs. inconsistent) was not significant, $p = .944$, indicating that the source of the (un)fairness – whether intentional or not – did not affect participants' behavior. Thus, results suggest that participants' behavior was not driven by a motivation to reciprocate an (un)fair counterpart, but rather reflected a mere reaction to being treated (un)fairly.

Lastly, since there was no effect of the Allocation condition, we collapsed across the Allocation conditions in our comparison of the prevalence of dishonest helping after being treated fairly and dishonest harming after being treated unfairly. A chi-squared analysis revealed that the proportion of dishonest helpers among those who received a fair amount (60.49%) was higher than the proportion of dishonest harmers among those who received an unfair amount (10.08%), $\chi^2(1) = 57.77$, $p < .001$, Cramer's $V = .537$. However, the rather high proportion of dishonest helpers among participants who received an unfair amount (48.73%) seems to suggest that people are rather likely to engage in dishonest helping, regardless of how they were treated (fairly or unfairly), and regardless of whether this treatment was intentional or not.

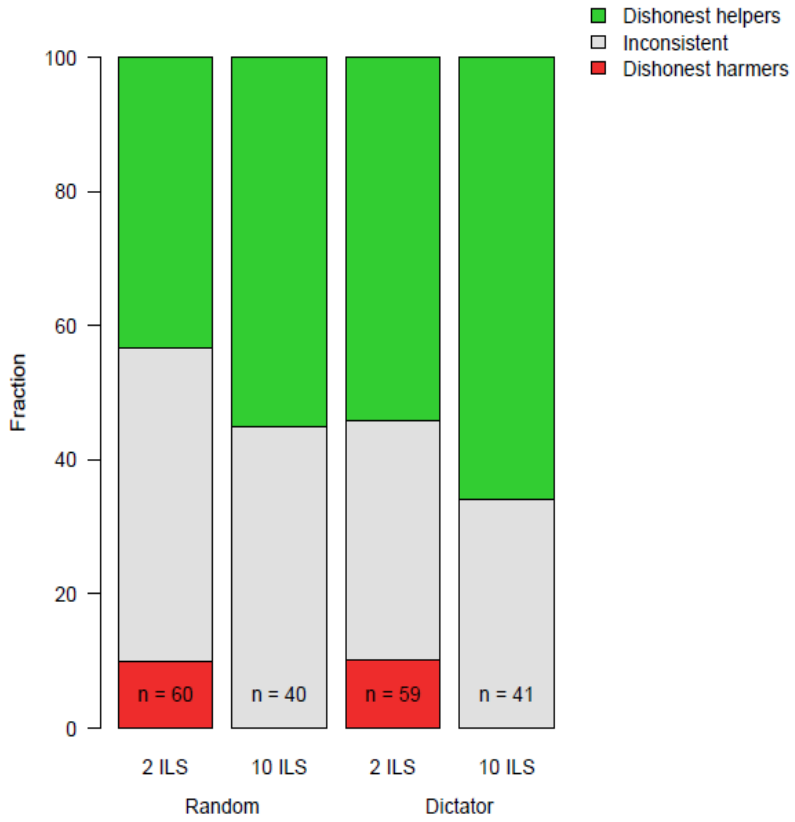

Figure 3.2. The fraction of behavioral types (dishonest helpers, dishonest harmers, and inconsistent), as a function of the amount participants received (2 ILS vs. 10 ILS) and the allocation condition (random vs. dictator). The *N*s of each group appear on the bar.

Emotions. As in Experiment 3.1 and 3.2, since emotions were measured after participants' behavior, we assess the association between dishonest reactions to (un)fairness and emotions and refrain from making causal claims. A series of ANOVA analyses with 2 (Behavior: dishonest helping after being treated fairly/dishonest harming after being treated unfairly vs. a different reaction to being treated (un)fairly) by 2 (Amount: unfair [2 ILS] vs. fair [10 ILS]) by 2 (Allocation: dictator vs. random) predicting gratitude, anger, and disappointment revealed a main effect for Amount on gratitude, anger, and disappointment. Participants who received a fair amount reported overall higher levels of gratitude (M = 4.97, SD = 1.30) than those who received an unfair amount (M = 3.71, SD = 1.35), $F(1, 191)$ = 68.37, $p < .001$, $\eta^2 = .264$. Further, those who received a fair amount were less angry (M = 1.25, SD = 0.98) than those who received an unfair amount (M = 1.99, SD = 1.68), $F(1, 187)$ = 52.60, $p < .001$, $\eta^2 = .220$. Similarly, those who received a fair amount were less disappointed (M = 1.26, SD = 0.83) than those who received an unfair amount (M = 2.23, SD = 1.96), $F(1, 190)$ = 57.79, $p < .001$, $\eta^2 = .233$.

As in Experiment 3.2, the Behavior × Amount interaction was significant for gratitude, $F(1, 191) = 18.64$, $p < .001$, $\eta^2 = .089$, anger: $F(1, 187) = 40.35$, $p < .001$, $\eta^2 = .177$, and disappointment, $F(1, 190) = 32.42$, $p < .001$, $\eta^2 = .014$. Lastly, the three-way interactions between Allocation, Amount, and Behavior were not significant for any emotion, ps > .090.

Assessing simple effects for the Behavior × Amount interaction, revealed that participants who received a fair amount reported similar levels of gratitude, anger, and disappointment regardless of whether they did or did not engage in dishonest helping. In contrast, those who received unfair amounts reported different levels of gratitude, anger, and disappointment depending on whether they did or did not dishonestly harm their counterpart. Among participants who received an unfair amount, those who engaged in dishonest harming reported that they were angrier, more disappointed, and less grateful than those who did not engage in dishonest harming, see Table 3.5 and SOM for detailed analyses.

Table 3.5. *Self-reported emotions in Experiment 3.3.*

	n	Gratitude	Anger	Disappointment
Unfair amounts (2 ILS)				
Dishonest harming	12	2.16 (1.20)	4.33 (2.42)	4.75 (2.30)
No dishonest harming	107	3.88 (1.26)	1.72 (1.34)	1.94 (1.71)
Difference		***	***	***
Fair amounts (10 ILS)				
Dishonest helping	49	5.13 (1.15)	1.06 (0.44)	1.14 (0.50)
No dishonest helping	31	4.72 (1.49)	1.55 (1.43)	1.45 (1.17)
Difference				

Note. Means (SDs) of the level of gratitude, anger, and disappointment 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 interactions with allocation (random vs. dictator) were not significant, the means reported here are collapsed across the allocation condition. Significance level: $^{***}p < .001$. Adjusting significance level for all the measures we collected (5 in total, see SOM), the new significance level is 0.05/5 = 0.01. $p < .01$ will be considered significant, thus all comparisons marked as *** remain significant.

Discussion

Replicating Experiment 3.2, in Experiment 3.3 we find that dishonest helping is prevalent. A meaningful proportion of individuals lied in order to help their counterparts, after

experiencing both fair and unfair treatment. It was only after experiencing unfair treatment that a rather small fraction of participants (~10%) engaged in dishonest harming. As in Experiment 3.2, after unfair treatment, dishonest harming was associated with higher levels of anger and disappointment, and lower gratitude than other reactions to unfair treatment.

Experiment 3.3 further allowed assessing whether recipients' dishonest helping and harming was driven by reciprocal motivation or by the mere feeling of being treated (un)fairly. Intriguingly, and contrary to our ex-ante prediction, participants responded to (un)fairness similarly when it was determined by a dictator, and when it was determined randomly. This behavioral pattern is consistent with participants' similar evaluation of fairness when the same monetary split was determined by a dictator and randomly.

One potential reason for this pattern of results might be the fact that completing the ambiguous die paradigm takes a rather long time – approximately 10 minutes. Prior work showed that a delay of around 10 minutes increases the acceptance rate of low monetary offers in an ultimatum game from ~20% to ~70% (Grimm & Mengel, 2011). That is, as time goes by, people are less likely to negatively reciprocate unfair offers by rejecting an offer and harming the offer maker and themselves. It thus might be the case that at the beginning of the task, participants in the dictator condition reacted differently than participants in the random allocation condition, but over time their motivation to reciprocate the (un)fair counterpart was diminished. To test this possibility, we conducted exploratory analyses in which we tested whether the Amount ×Allocation interaction predicted participants' reports, focusing on the first trials of the task. Our results, however, revealed that also when restricting our analyses to the first trials (first trial, first 5 trials, and first 10 trials), the Amount × Allocation interaction was not significant, *p*s > .542, see SOM for full analyses. We thus conclude that it is the mere (un)fair treatment and not reciprocal motivation that drove the behavior obtained here.

General discussion

People are treated in fair and unfair ways all the time. At times, they react to such (un)fair treatments by breaking the rules and lying. Here we assess the prevalence of dishonest harming and helping after (un)fair treatment. Across three financially incentivized experiments we find that overall, people are likely to engage in dishonest behavior aimed at helping others. Dishonest helping seems to be a default behavior that occurs both when people experience fair treatment, as well as when people do not experience any treatment at all. Only when experiencing unfair treatment, do some people change their default behavior and start engaging in dishonest harming. Thus, consistent with the notion that people are more sensitive

to negative, compared to positive events (Baumeister et al., 2001; Klein & Epley, 2014; Kube et al., 2006), we conclude that also when engaging in dishonest behavior aimed at affecting others' pay, people react stronger to being treated unfairly, compared to being treated fairly.

Assessing people's emotions further revealed that dishonest reactions to fairness were not associated with their emotions, whereas dishonest reactions to unfair treatment were. In particular, among participants who experienced unfair treatment, those who engaged in dishonest harming also reported higher level of anger and disappointment, and lower levels of gratitude compared to those who did not engaged in dishonest harming. However, among participants who experienced fair treatment, there was no association between people's level of gratitude, anger, and disappointment and their dishonest helping behavior. This set of results further points toward higher sensitivity to unfair, compared to fair treatment. Since in all three experiments we measured participants' emotions at the end of the task we interpret these results with caution and refrain from making any causal inference. It might be that experiencing unfair treatment makes some people feel angrier, more disappointed, and less grateful, which in turn pushes these individuals to engage in dishonest harming. On the contrary, it might be that some individuals react to unfair treatment by engaging in dishonest harming, and in turn rationalize their behavior by stating that they felt angry, disappointed, and ungrateful.

Interestingly, we find a rather high prevalence of dishonest helping among participants who experienced unfair treatment. In fact, in Experiment 3.2 and 3.3, where we could assess dishonest behavior at an individual level, dishonest helping was a more common reaction to unfair treatment than dishonest harming. Among those who received unfair treatment in Experiment 3.2, there were 36.58% dishonest helpers and 24.39% dishonest harmers. Similarly, among those who received unfair treatment in Experiment 3.3, there were 48.73% dishonest helpers and only 10.08% dishonest harmers. This finding is in line with the fact that generally people do not like to harm others (Baron, 1995), and view dishonest helping as a rather ethical and noble action (Gino & Pierce, 2010a; Levine & Schweitzer, 2014). While dishonest helping can be perceived in a positive light and be viewed as a prosocial behavior, it is important to keep in mind that dishonest helping often comes at a cost to third parties (e.g., experimental budget, reduced trust), and can be a fertile ground for developing corrupt collaboration (Weisel & Shalvi, 2015; Gross, Leib, Offerman & Shalvi, 2018; Soraperra et al., 2017).

Future directions and limitations

In Experiment 3.3, participants' behavior did not differ when the (un)fair allocation was determined by a dictator or by a random allocation, suggesting that the mere feeling of being treated (un)fairly, rather than the motivation to reciprocate the (un)fair counterpart accounted for the results obtained here. Interestingly, prior work did find that people react to (un)fairness differently when it was intentional versus not (Falk, Fehr, & Fischbacher, 2008, Falk & Fischbacher, 2006; Offerman, 2002). We see two possibilities to the potential difference between our work and prior findings. One possibility is that participants did not believe the information they received regarding how the monetary allocation was determined (by a dictator versus randomly). If that is the case, we should not observe differences between the conditions. We believe this possibility is unlikely as the experiment was run in a behavioral economics lab, which strictly maintains a non-deception policy, a fact that participants are fully aware of.

A second possibility is that reactions to (un)fairness that entail dishonest behavior lead to a different behavioral pattern that reactions to (un)fairness that do not entail dishonest behavior. It might be the case that for those who engaged in dishonest helping and harming, the mere feeling of being treated (un)fairly was sufficient to push them to lie. The source of the (un)fairness, whether it was intentional or not, might not be needed as an additional reason for those who have already decided to lie. For those who have decided not to engage in dishonest behavior, the intentionality of the (un)fair treatment might not be a sufficient push to break ethical rules and lie. A promising avenue for future work can be to assess the role intentions play in dishonest (and honest) reactions to other experiences such as cooperation, charity giving, and being the victim of deception oneself.

In experiments 3.1 and 3.3, the highest amount dictators could give to their counterparts was 50% of their endowment. In Experiment 3.2, where we did not restrict dictators' choices, the highest allocation dictators gave was also 50% of their endowment (with an exception of two dictators who gave 100% of their endowment). A recent meta-analysis on the dictator game reported that out of 20,813 dictator allocation decisions, the vast majority (86.94%) gave up to 50% of their endowment (Engel, 2011). Further, the two most common allocations were 0% of the endowment (given by 36.11% of the dictators) and 50% of the endowment (given by 16.74% of the dictators). As such, results obtained here provide insights about people's reactions to the most common levels of (un)fair treatments. However, since we did not manipulate dictators' decisions, but measured them, we did not capture

dishonest reactions to extremally generous dictator allocations, such as 100% of the endowment. It might be the case that when people experience extreme generosity, they react with even higher levels of dishonest helping. It is thus possible that compared to receiving 50% of a dictator's endowment, people's dishonest helping after receiving 100% of the endowment is more prevalent than their dishonest harming after receiving 0% of the endowment. Assessing people's dishonest reactions to extremely generous allocations and comparing it to reactions to extremely ungenerous allocations is yet another interesting avenue for future work.

In this work, we assess dishonest reactions to behavior that affects the individual directly. In our settings, participants were the ones who received an (un)fair amount of money and then reacted to it. Assessing how people react to first-hand experiences is important. However, people also witness behaviors that are directed toward others and react to them, even when they are not directly affected by them. One prime example is the existence of third-party punishment (Fehr & Fischbacher, 2004; Ule, Schram, Riedl, & Cason, 2009), in which people sacrifice their own payoffs in order to punish uncooperative others in an attempt to enforce the social norm of cooperation. Another interesting directions for future work is to test whether people are likely to lie to punish uncooperative others and reward cooperative others, even when they are not affected by the (un)cooperative act themselves, and whether incidental anger (Yip & Schweitzer, 2016) drives such behavior.

Lastly, the prevalence of dishonest helping and harming may vary when they are achieved by lies of omission versus lies of commission. Lies of omission, where individuals refrain from telling the truth, have been found to be more common than lies of commission, where individuals need to actively lie (Mazar & Hawkins, 2015). This is because lying by withholding information is perceived as more legitimate and justifiable than lying by providing false information (Pittarello, Rubaltelli, & Motro, 2016, see also Spranca, Minsk, & Baron, 1991). Further, people prefer to endure harm caused by inaction than harm caused by action, even when the harm caused by action is smaller (Baron & Ritov, 2004). It may thus be the case that dishonest harming will be more prevalent when doing so requires individuals to refrain from telling the truth, compared to actively lie. Since in many situations one can help or harm others by not telling the truth, rather than actively lying (e.g., not reporting a colleague's rule violation), assessing the role of active vs. passive lies in shaping dishonest reactions to (un)fair treatment seems like an important avenue for future exploration.

Conclusion

People experience (un)fair treatment and react to it regularly. Here we assess the extent to which individuals are willing to engage in dishonest helping and harming behavior after they experience (un)fair treatment. We find that in general, dishonest helping is a very common and robust behavior. Individuals lie to help others – regardless of their prior experience. We found evidence for dishonest helping among those who experience fair, unfair, and no prior treatment. Dishonest harming, on the other hand, was less prevalent. Only after unfair treatment did some individuals engage in dishonest harming. To prevent dishonesty from emerging and spreading, it is important to craft environments in which people are encouraged to treated each other in a fair manner, and can help each other in honest, ethical ways.