

Classification: Social Sciences, Psychological and Cognitive Sciences

Corrupt Collaboration

Ori Weisel¹ & Shaul Shalvi²

¹ Centre for Decision Research and Experimental Economics, University of Nottingham, Nottingham, NG7 2RD, United Kingdom.

² Department of Psychology, Ben-Gurion University of the Negev, Be'er-Sheva 84105, Israel.

Corresponding authors:

Ori Weisel; Centre for Decision Research and Experimental Economics, University of Nottingham, Nottingham, NG7 2RD, United Kingdom; +44 (0) 115 9515458; ori.weisel@nottingham.ac.uk.

Keywords: Cooperation; Corruption; Decision Making; Behavioral Ethics; Behavioral Economics

Abstract

Cooperation is essential for completing tasks that individuals cannot accomplish alone. Whereas the benefits of cooperation are clear, little is known about its possible negative aspects. Introducing a novel sequential dyadic die-rolling paradigm, we show that collaborative settings provide fertile ground for the emergence of corruption. In the main experimental treatment the outcomes of the two players are perfectly aligned. Player A privately rolls a die, reports the result to player B, who then privately rolls and reports the result as well. Both players are paid the value of the reports if, and only if, they are identical (e.g., if both report 6, each earns €6). Since rolls are truly private, players can inflate their profit by misreporting the actual outcomes. Indeed, the proportion of reported doubles was 489% higher than the expected proportion assuming honesty, 48% higher than when individuals rolled and reported alone, and 96% higher than when lies only benefited the other player. Breaking the alignment in payoffs between player A and player B reduced the extent of brazen lying. Despite player B's central role in determining whether a double was reported, modifying the incentive structure of either player A or player B had nearly identical effects on the frequency of reported doubles. Our results highlight the role of collaboration—particularly on equal terms—in shaping corruption. These findings fit a functional perspective on morality. when facing opposing moral sentiments—to be honest vs. to join forces in collaboration—people often opt for engaging in corrupt collaboration.

Significance Statement

Recent financial scandals highlight the devastating consequences of corruption. While much is known about individual immoral behavior, little is known about the roots of “corrupt collaboration”. In a novel experimental paradigm people could adhere to one of two competing moral norms: collaborate vs. be honest. Whereas collaborative settings may boost honesty due to increased observability, accountability, and reluctance to force others to become accomplices, we show that collaboration, particularly on equal terms, is inductive to the emergence of corruption. When partners' profits are not aligned, or when individuals complete a comparable task alone, corruption levels drop. These findings reveal a dark side of collaboration, suggesting that human cooperative tendencies, and not merely greed, take part in shaping corruption.

\body

Corrupt Collaboration

Humans are an exceptionally cooperative species. We cooperate in groups which extend beyond the boundaries of genetic kinship even when reputational gains are unlikely. Such cooperative tendencies are at least partly driven by deeply ingrained moral sentiments reflected in a genuine concern for the well-being of others (1), and allow people to build meaningful relationships (2, 3), develop trust (4, 5), achieve mutually beneficial outcomes (6–8), and strengthen bonds with in-group members (9, 10). Furthermore, reciprocating others' cooperative acts is essential for establishing long-term cooperation (11–13). Clearly, establishing sustainable cooperative relationships can set successful individuals and groups apart from less successful ones (1).

While the benefits of cooperation are clear, little is known about its possible negative aspects (14). Our interest is in cases in which the collaborative effort of individuals working together necessarily and directly entails the violation of moral rules (here: lying), at a possible cost to the larger group, or the organization, to which they belong. In such cases there is a tension between two moral obligations: to tell the truth, and to collaborate. Are people prepared to accept the costs associated with violating moral rules (i.e., getting caught and punished, as well as the psychological costs associated with lying) in order to establish collaborative relations with others?

The answer is not trivial. On the one hand, as collaboration involves increased observability and accountability, reputational concerns may limit the willingness to violate moral rules. People may also shy away from lying in collaborative settings to avoid imposing undeserved, and potentially unwanted, profits on their partners, forcing them to become accomplices or “partners in crime”. On the other hand, a number of findings support the view that collaboration might have a liberating effect, freeing people to behave unethically. People lie

more when it improves not only their own (14–16), but also others' outcomes (18–22); when their lies benefit a cause or another person they care about (23); and group-serving dishonesty is modulated by oxytocin, a social bonding hormone (24). The idea that collaboration will increase dishonest behavior is in line with the functionalist approach to morality, which prescribes that people treat morality in a flexible manner, judging the same act as illegitimate or immoral in some cases, but legitimate or even moral in other cases, e.g., when it profits one's group members (25, 26).

Some collaborative settings may be more prone to corrupt behavior than others. In particular, corruption may emerge more readily when parties share profits equally. Indeed, equality is perhaps the most prevalent fairness norm (27–29); people are willing to pay a cost to restore it (30) and to punish others who violate it (31). Here, we conjecture that corrupt collaboration—the attainment of personal profits by joint immoral acts—would be (i) particularly prevalent when both interaction partners equally share the profits generated by dishonest acts, and (ii) more frequent than individual dishonest behavior in a comparable setting.

We experimentally examined corrupt collaboration using a novel sequential dyadic die-rolling paradigm. In seven treatments participants (N=280; 20 dyads per treatment; data collected in Germany) were paired and assigned to role A or B. In an additional individual treatment (N=36) the same person acted in both roles (See SI Appendix). As depicted in Fig. 1, each participant privately rolled a die and reported the outcome. Player A rolled a standard six-sided die first, and reported the outcome by typing a number on the computer. Subsequently, Player B was informed of A's reported number, and proceeded to roll a die and report the outcome as well. The information about B's reported outcome was then shared with A. The interaction was repeated for twenty trials.

Die rolls were truly private, allowing participants to misreport the actual outcomes without any fear of being caught (17, 32). Such privacy is key to our design, as it reflects real-life situations in which cutting moral corners is difficult, if at all possible, to detect. Participants' reported outcomes determined their payoff, according to rules that were explained in advance, and varied between the treatments (See *Materials and Methods* and SI Appendix).

In the main Aligned Outcomes treatment, participants earned money only if both A and B reported identical numbers (i.e., a 'double'). In these cases each of them earned the reported numbers in Euros (e.g., if both reported 4 each received €4). If the reports were not identical (i.e., not a double), neither A nor B earned any money. Reporting truthfully, the probability of rolling a double in a single trial is 1/6. The expected number of doubles reported by each pair in 20 trials is thus 3.33 (16.7%), and the mean report (of both A and B) is expected to be 3.5 ($[1+2+3+4+5+6]/6$). Panel A in Fig. 2 shows a simulation of outcomes assuming such truthful behavior. As can be seen in panel B in Fig. 2, however, reports in Aligned Outcomes were far from truthful (see SI Appendix, Fig. S10, for dyad-level figures). The twenty dyads reported 16.30 (81.5%) doubles on average, almost 5 times as much as the expected 3.33 assuming honesty (Wilcoxon Signed Rank: $U=210$, $P<0.001$; each dyad is a single observation). Corruption was not limited to inflation in the number of doubles. The mean reports of both A (5.02) and B (4.92) were significantly higher than the expected 3.5 (A: $U=208.5$, $P<0.001$; B: $U=171$, $P<0.001$), demonstrating that both parties—and not only player B—were dishonest.

The organizational costs associated with lying are represented by the experimental budget, as common in experimental work studying dishonesty (16, 27, 28). In the current setting the behavior of both players A and B affects these costs, but the situation is not symmetrical. Since player A can (only) *set the stage* by inflating his/her reports, but player B is the one

who can *get the job done* by matching A's report, the substantial organizational costs of corrupt behavior in this setting largely depend on player B. A dishonest player A can (only) increase the cost of reported doubles less than two-fold, from an expected value of 3.5 (for a perfectly honest A) to a maximum of 6 (for a completely brazen A, consistently reporting 6's). In contrast, when B lies in a totally brazen manner (i.e., always reports a double), the expected one-in-six likelihood of a double (assuming honesty) is shifted to a certain six-in-six (assuming full dishonesty), a six-fold increase. In Aligned Outcomes 25% of A-players and 50% of B-players exhibited such brazen behavior (see SI Appendix, Fig. S10).

To examine whether the degree to which incentives are aligned is related to the frequency of lying, we misaligned the incentives of A and B in a series of control treatments. Given B's central role, we first focus on treatments that vary B's incentives. A's incentives were kept as in Aligned Outcomes. Treatments B-high and B-low removed B's interest in the *value* of the double by having B earn a fixed amount— €6 or €1, respectively—if a double was reported, and nothing otherwise. Treatment B-fixed removed B's interest in *reporting* a double by having B earn a fixed amount of €1 regardless of whether a double was reported or not.

We expected that increasing B's incentive to lie (B-high) will result in more reported doubles; decreasing B's incentive to lie (B-low) will result in less reported doubles; and removing B's incentive to lie altogether (B-fixed) will lower the number of reported doubles even more. Possibly due to the already high levels in Aligned Outcomes, the number of reported doubles in B-high (15.3 doubles per dyad) was a bit lower—but not significantly different—than in Aligned Outcomes (Mann-Whitney: $U=165.5$, $P=0.34$). As expected the number of reported doubles was significantly reduced in B-low (12.2; $U=120$, $P=.028$), and even more so in B-fixed (8.3; $U=64.5$, $P<0.001$; see Fig. 3).

The number of doubles in B-fixed was still more than twice the 3.33 that would be expected assuming honest behavior ($U=185$, $P=.003$), demonstrating that B-players are willing to lie even when only A-players profit from it. Nevertheless, B-fixed is promising for an organization seeking to reduce corruption costs by providing an alternative to Aligned Outcomes as a way to compensate employees. In Aligned Outcomes each player earned €4.28 per trial, for a total of €8.56 in organizational costs. Considering that the expected cost, assuming that both players A and B are honest, is €1.16 ($1/6 \times €3.5 = €0.58$ for each player), this represents a ‘corruption excess’ of €7.4. In B-fixed B players earned exactly €1 per trial, and A players earned only €1.80 per trial (there were less doubles, and the average value of the doubles was lower), for a total of €2.80 in organizational costs, or a ‘corruption excess’ of only €1.64, a 78% reduction relative to Aligned Outcomes. Clearly, B-fixed is superior to Aligned Outcomes from the organization’s perspective.

Given that B’s selfish incentive to report a double was removed in B-fixed, it is perhaps not surprising that the number of reported doubles was reduced relative to Aligned Outcomes. That notwithstanding, B-fixed—at least at face value—is attractive from B’s perspective as compared to the Aligned Outcomes setting. Consider a potential employee who is negotiating employment terms, and is given a choice between a self-reported performance-based salary, similar to Aligned Outcomes, and a fixed salary, similar to B-fixed. In this initial negotiation phase, the potential profits the employee may reap by engaging in corrupt behavior in Aligned Outcomes are not officially ‘on the table’. The only thing to consider is the expected value of the two settings assuming honest behavior; otherwise, the employee would be signaling an intention to behave dishonestly in the future, reasonably terminating the organization’s willingness to offer this person a job altogether. The expected payoff—assuming honesty—from a single trial under Aligned Outcomes is €0.58 (see above); a guaranteed €1 per trial is clearly a better deal, unless an employee has dishonest intentions

and is willing to communicate those to the employer by rejecting a 72% payment increase (from €0.58 to €1). Offering B-fixed as an alternative compensation program to Aligned Outcomes not only reduces costs, but also provides the organization with a useful screening tool to detect people with dishonest intentions.

In some cases people work alone. Do people lie more when collaborating than when working alone? Additional participants (N=36) played the role of both A and B (i.e., each participant rolled and reported twice). Individuals reported 11.0 doubles (54.9%) on average, significantly less than dyads in Aligned Outcomes (Mann-Whitney: $U=187$, $P=0.003$), showing that collaborative settings indeed liberate people to lie more than when they work alone. This finding suggests that organizations may be paying a (corruption) premium for having their employees team up and work together. The modest dishonesty estimates observed in past work (17, 32, 33) may have been conservative in comparison with settings in which people work in collaboration.

Player A's incentive to report a double was varied in three additional control treatments—A-high, A-low, and A-fixed—that mirror the variations to B's incentives in B-high, B-low, and B-fixed. B's incentives were kept constant as in Aligned Outcomes. Would these changes to A's incentive structure affect the frequency of reported doubles to the same degree as (the same) changes to B's incentives? Since B is the one determining whether a double is reported (or not), changes to A's incentives should have a smaller impact—if any—on the amount of reported doubles. Surprisingly, however, the impact of changes in A's incentives on the likelihood to report a double was nearly identical to that of changes in B's incentives (A-low vs. B-low: $U=176$, $P=0.52$; A-high vs. B-high: $U=197$, $P=0.95$; A-fixed vs. B-fixed: $U=160.5$, $P=0.29$; see Fig. 3).

An important difference between the Aligned Outcomes setting and the other (control) treatments is the propensity of B players to behave in a brazen manner (i.e., to report a double in each and every trial). In many situations, people tend to avoid brazen immoral behavior, offsetting the psychological costs of immoral behavior by engaging in moral behavior as well (32, 33), a pattern observed in our control treatments. In Aligned Outcomes, however, the behavior of B-players took a rather brazen form. While the likelihood of honestly rolling a double in each and every one of the twenty trials is 2.7×10^{-16} , this occurred in ten out of twenty cases in the Aligned Outcomes setting. Half of the B-players in Aligned Outcomes behaved in a totally brazen manner, making no effort at all to even remotely suggest truthful reporting. See Fig. 4 for examples of such brazen behavior, which was more common in Aligned Outcomes when compared to all other dyadic (15%; Fisher's exact test: $P=0.001$) and individual (19%; $P=0.03$) settings (see Fig. 3). Collaboration on equal terms, it seems, liberates people to turn corrupt by reducing the psychological costs associated with brazen lying.

The finding that changes in A's incentives affect the number of reported doubles suggests that B players are sensitive to the incentives, and possibly behavior, of A players. To test this notion, we classified each player as either brazen (i.e., A-players reporting "six" twenty times; B players reporting twenty doubles) or not, and compared the proportion of brazen B players who were matched with brazen vs. non-brazen A players. If players A and B have no effect on each other's behavior, the likelihood of B being brazen should be independent of whether A is brazen or not. If however, players A and B do affect each other's behavior, the likelihood of B being brazen should be higher when A is brazen as well. This is indeed what we found. In the Aligned Outcomes treatment, 100% of B's were brazen when A was brazen as well (5 of 5 cases). When A was not brazen the proportion of brazen B's was only 33.33% (5 of 15 cases; $\chi^2(1) = 6.67, P < .01$). This pattern was attenuated in all other treatments where

the proportion of brazen B's when A was brazen as well was 36.36% (4 of 11 cases), as compared to 12.84% (14 of 109 cases) when A was not brazen ($\chi^2(4) = 26.50$, $P < .001$; the result holds also when excluding the A-fixed and B-fixed treatments).

The fact that B's were more likely to turn brazen when A was brazen suggests that people perceive their counterpart's brazen behavior as an 'invitation to lie' (i.e., a signal) that justifies their own corrupt behavior. It is especially interesting that this association between the two party's brazenness is stronger in Aligned Outcomes compared to the other treatments. When incentives are perfectly aligned, one party's brazen behavior is more likely to be picked up and serve as justification for the brazen behavior of the partner. More generally, this result suggests that (corrupt) signals are easier to interpret when outcomes are aligned.

A possible concern is that our findings are idiosyncratic to particular characteristics of the experiment, namely the location and the stakes we used. To assess the robustness of our results we conducted an additional experiment (N=236) in the United Kingdom, with three variations of the original Aligned Outcomes and B-fixed treatments: Replication, Multiplication, and Addition. The Replication treatments (N=80, 20 dyads per treatment) directly replicated the original Aligned Outcomes and B-Fixed treatments (the only difference was using British Pounds rather than Euros). In the Multiplication treatments all payoffs were doubled (e.g., a double of 4's yielded £8). In the Addition treatments £2 were added to all payoffs (e.g., the payoff for a double of 4's was £6).

The results are displayed in Fig. 5. As in the original experiment, in all treatments (Replication, Multiplication, Addition) more doubles were reported in Aligned Outcomes than in B-fixed ($P < .001$, $P < .001$, $P = .03$, respectively). Despite the different subject pool, the absolute levels of dishonest behavior in the Replication treatments were very similar to those in the original experiment (Aligned Outcomes: $U = 167.5$, $P = .37$; B-fixed: $U = 145$, $P = .14$).

Furthermore, multiplying the payoffs by two did not affect the number of reported doubles (Aligned Outcomes: $U=178.5$, $P=.98$; B-fixed: $U=177$; $P=.54$; comparison between Replication and Multiplication). These results are consistent with past work in individual settings, which found that increasing the stakes does not affect lying rates (17, 33). The Replication and Multiplication settings are also similar in that in both we observed more brazen behavior in Aligned Outcomes than in B-fixed (30% vs. 5% and 33% vs. 0%, respectively; both $P's < .05$), again replicating the results obtained in the main experiment. Finally, further replicating the original results, in both the Replication and Multiplication versions of Aligned Outcomes there were more brazen B players when A was brazen (5 of 7; 71%), as compared to when A was not brazen (7 of 33; 21%; $\chi^2(1) = 6.94$, $P=.008$). In B-fixed there was only one brazen A player, not allowing for a meaningful analysis of the way A's brazen behavior is related to B's brazen behavior.

In the Addition setting the pattern was somewhat different. Relative to the Replication setting, there was less lying in Aligned Outcomes (the difference is marginally significant; $U=136$, $P=.08$), but not in B-Fixed ($U=157.5$, $P=.25$). Further, Aligned Outcomes and B-fixed did not differ in the rates of brazen behavior (15% vs. 10%, $P=.63$). Since only one A player was brazen in the Addition setting (Aligned Outcomes and B-Fixed together), it was not possible to analyze whether A's brazen behavior was related to B's brazen behavior. The results from the Addition setting suggest that even if the absolute incentive to lie and report a double remains unchanged, earning at least something rather than nothing, in case of failure to report a double, can reduce the likelihood of brazen lying and may limit the emergence of corrupt collaboration. From the point of view of an organization seeking to reduce corrupt behavior, assuring a decent base salary that does not depend on performance can reduce the likelihood that its employees engage in brazen lying.

The current work reveals a dark side of cooperation: corrupt collaboration. A collaborative setting led people to engage in excessive dishonest behavior. The highest levels of corrupt collaboration occurred when the profits of both parties were perfectly aligned, and were reduced when either player's incentive to lie was decreased or removed. These results suggest that acts of collaboration, especially on equal terms, constitute 'moral currencies' in themselves, which can offset the moral costs associated with lying. Paradoxically, the corrupt corporate culture and brazen immoral conduct at the roots of recent financial scandals (34) are possibly driven not only by greed, but also by cooperative tendencies and aligned incentives. In conclusion, when seeking to promote collaboration in our organizations and society, we should take note that in certain circumstances cooperation should be monitored, rather than encouraged unambiguously.

Materials and Methods

In the main experiment three hundred and sixteen undergraduate students (56% females; $M_{\text{age}}=24.39$, $SD_{\text{age}}=4.84$) were recruited using ORSEE (35), and took part in 11 experimental sessions, each consisting of between 20 and 32 participants, and lasting less than an hour on average. We used a within-session design, meaning that all eight treatments (Aligned Outcomes, B-high, B-low, B-fixed, A-high, A-low, A-fixed, Individuals) took place simultaneously in the same session. Participants were given general instructions on paper, and treatment-specific instructions were presented on-screen.

The general instructions, which were read out loud by an experimenter, informed participants that they will engage in a study composed of three stages. Each stage was explained separately once the previous stage was completed. Participants learned that they will earn money based on their performance in each of the three stages. In the first stage, participants learned that they will engage in a die-rolling task in which the amount of money they will

earn depends on the results of two die rolls. Following the general instructions, participants received on-screen treatment-specific information about the rules determining their payoff (according to the treatment they are assigned to; see “Treatments” below). To ensure proper understanding, participants were instructed to roll the die a couple of times and type in the results. They received feedback about the payoffs associated with these rolls, according to the relevant rules. After indicating they understood the rules, participants were paired (except those in the Individual treatment), and played one period of the die-rolling task for actual payoffs. Upon completion of the first stage, participants were informed that in the second stage they will engage in up to thirty additional periods with a different partner, and that they will be paid according to the results of one of these periods, which will be randomly selected. In reality there were always twenty additional periods.

We varied participants' payoffs in eight between-subjects experimental treatments (N=40 in each of the two-person treatments; N=36 in the individual treatment). In the main, Aligned Outcomes, treatment, if both A and B reported the same number, each earned that number in Euros (i.e., if both reported 1, each earns €1; if both reported 2, each earns €2; etc.). In case they reported different numbers, they earned nothing. In six control treatments the payoff of either A or B was varied. In treatment A-high (B-high) A (B) earned €6 if a double was reported (regardless of the value of the double) whereas B (A) earned the value of the double as in Aligned Outcomes. In treatment A-low A (B) earned €1 if a double was reported (regardless of the value of the double) whereas B (A) earned the value of the double as in Aligned Outcomes. In treatment A-fixed (B-fixed) A (B) earned a fixed €1 in each period, regardless of whether a double was reported or not, whereas B (A) earned the value of the double as in Aligned Outcomes. In an Individual treatment, the same person rolled and reported twice, and earned the value of the double if a double was reported, nothing otherwise.

Finally, in the third stage we assessed participants' Social Value Orientation using the SVO slider measure (See SI Appendix, section 1.1.7). The measure assesses the magnitude of people's concern to others' outcomes. Each participant made a series of 15 choices between nine resource allocation options between self and another participant (i.e., in each choice there were nine available allocations to choose from). Payment for this part was based on one of these choices (randomly determined). The 15 choices were aggregated to determine a unique value for each participant, expressed as an angle on a self/other two-dimensional space. The angle was used to classify participants as competitive, individualistic, pro-social, or altruistic. We followed (36) to determine the borders between the four SVO types.

For the robustness-experiment, two hundred and thirty six undergraduate students (65% females; $M_{\text{age}}=21.32$, $SD_{\text{age}}=3.16$) were recruited using ORSEE (35), and took part in 9 experimental sessions, each consisting of 24 or 28 participants, and lasting less than an hour on average. We used a within-session design, meaning that the Aligned Outcomes and B-fixed treatments within each setting (Replication, Multiplication, Addition) took place simultaneously in the same session. Participants were given general instructions on paper, and treatment-specific instructions were presented on-screen. Payment in these sessions was given in British Pound Sterling, as they were conducted at the University of Nottingham.

Each setting—Replication, Multiplication, and Addition—included two treatments: Aligned Outcomes and B-fixed. In the Replication setting the payoffs were exactly as in the original experiment, except that they were given in £'s instead of €'s. In the Multiplication setting the payoffs were exactly double the amounts paid in the Replication setting. In the Addition setting each player received an additional fixed amount of £2 per trial (relative to the Replication), which was added to the earnings from the reported outcomes.

Acknowledgements. We thank Abigail Barr, Benjamin Beranek, Simon Gächter, Georgia Michailidou, Lucas Molleman, Marieke Roskes, Jonathan Schultz, Till Weber, and three anonymous reviewers for their helpful comments. The research was funded by the Max Planck Society, the European Research Council (ERC-AdG 295707 COOPERATION), and the European Union’s Seventh Framework Program (FP7/2007-2013; REA 333745). The authors declare no competing interests. Author contribution: OW and SS designed the experiments, OW programmed the materials, collected and analyzed the data, OW and SS wrote the paper.

References:

1. Bowles S, Gintis H (2011) *A cooperative species: Human reciprocity and its evolution* (Princeton University Press).
2. Baumeister RF, Leary MR (1995) The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychol Bull* 117(3):497.
3. Kameda T, Takezawa M, Hastie R (2005) Where do social norms come from? The example of communal sharing. *Curr Dir Psychol Sci* 14(6):331–334.
4. Kramer RM (1999) Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annu Rev Psychol* 50(1):569–598.
5. Rempel JK, Holmes JG, Zanna MP (1985) Trust in close relationships. *J Pers Soc Psychol* 49(1):95.
6. Bazerman MH, Curhan JR, Moore DA, Valley KL (2000) Negotiation. *Annu Rev Psychol* 51(1):279–314.
7. Murnighan JK, Murnighan JK (1992) *Bargaining games: A new approach to strategic thinking in negotiations* (W. Morrow).
8. Rusbult CE, Van Lange PA (2003) Interdependence, interaction, and relationships. *Annu Rev Psychol* 54(1):351–375.
9. De Dreu CKW, et al. (2010) The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. *Science* 328(5984):1408–1411.
10. Efferson C, Lalive R, Fehr E (2008) The coevolution of cultural groups and ingroup favoritism. *Science* 321(5897):1844–1849.
11. Axelrod R, Hamilton WD (1981) The evolution of cooperation. *Science* 211(4489):1390–1396.
12. Dawes RM (1980) Social dilemmas. *Annu Rev Psychol* 31:169–193.
13. Rand DG, Dreber A, Ellingsen T, Fudenberg D, Nowak MA (2009) Positive interactions promote public cooperation. *Science* 325(5945):1272–1275.
14. Van Lange PA, Joireman JA (2008) How we can promote behavior that serves all of us in the future. *Soc Issues Policy Rev* 2(1):127–157.

15. Schweitzer ME, Ordóñez L, Douma B (2004) Goal setting as a motivator of unethical behavior. *Acad Manage J* 47(3):422–432.
16. Coricelli G, Joffily M, Montmarquette C, Villeval MC (2010) Cheating, emotions, and rationality: an experiment on tax evasion. *Exp Econ* 13(2):226–247.
17. Fischbacher U, Föllmi-Heusi F (2013) Lies in disguise—an experimental study on cheating. *J Eur Econ Assoc* 11(3):525–547.
18. Wiltermuth SS (2011) Cheating more when the spoils are split. *Organ Behav Hum Decis Process* 115(2):157–168.
19. Gino F, Ayal S, Ariely D (2013) Self-serving altruism? The lure of unethical actions that benefit others. *J Econ Behav Organ* 93:285–292.
20. Conrads J, Irlenbusch B, Rilke RM, Walkowitz G (2013) Lying and team incentives. *J Econ Psychol* 34:1–7.
21. Cohen TR, Gunia BC, Kim-Jun SY, Murnighan JK (2009) Do groups lie more than individuals? Honesty and deception as a function of strategic self-interest. *J Exp Soc Psychol* 45(6):1321–1324.
22. Gneezy U (2005) Deception: The role of consequences. *Am Econ Rev*:384–394.
23. Lewis A, et al. (2012) Drawing the line somewhere: An experimental study of moral compromise. *J Econ Psychol* 33(4):718–725.
24. Shalvi S, De Dreu CKW (2014) Oxytocin promotes group-serving dishonesty. *Proc Natl Acad Sci* 111(15):5503–5507.
25. Haidt J (2007) The new synthesis in moral psychology. *Science* 316(5827):998–1002.
26. Greene J (2014) *Moral tribes: emotion, reason and the gap between us and them* (Atlantic Books Ltd).
27. Fehr E, Schmidt KM (1999) A theory of fairness, competition, and cooperation. *Q J Econ*:817–868.
28. Bolton GE, Ockenfels A (2000) ERC: A theory of equity, reciprocity, and competition. *Am Econ Rev*:166–193.
29. Andreoni J, Bernheim BD (2009) Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica* 77(5):1607–1636.
30. Dawes CT, Fowler JH, Johnson T, McElreath R, Smirnov O (2007) Egalitarian motives in humans. *Nature* 446(7137):794–796.
31. Fehr E, Fischbacher U (2004) Third-party punishment and social norms. *Evol Hum Behav* 25(2):63–87.
32. Shalvi S, Dana J, Handgraaf MJ, De Dreu CKW (2011) Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organ Behav Hum Decis Process* 115(2):181–190.
33. Mazar N, Amir O, Ariely D (2008) The dishonesty of honest people: A theory of self-concept maintenance. *J Mark Res* 45(6):633–644.
34. Cohn A, Fehr E, Marechal MA (2014) Business culture and dishonesty in the banking industry. *Nature* advance online publication. Available at: <http://dx.doi.org/10.1038/nature13977>.
35. Greiner B (2004) An online recruitment system for economic experiments. *Forschung Und Wissenschaftliches Rechnen (GWVG Bericht 63)*, eds Kremer K, Macho V (Gesellschaft für Wissenschaftliche Datenverarbeitung, Göttingen), pp 79–93.

36. Murphy RO, Ackermann KA, Handgraaf MJ (2011) Measuring social value orientation. *Judgm Decis Mak* 6(8):771–781.

Figures:

Fig. 1. Procedure for Aligned Outcomes (see SI Appendix for other treatments)





<p>Procedure</p> <ol style="list-style-type: none">1. Player A privately rolls a die and reports the outcome2. Player B learns about A's report3. Player B privately rolls a die and reports the outcome4. Player A learns about B's report	<p>Example 1 - Double</p> <p>Player A reported </p> <p>Player B reported </p> <p>Payment: A gets €5, B gets €5</p>
<p>Payment (Aligned Outcomes)</p> <p>If the reported outcomes are equal each player receives the reported amount in Euros.</p> <p>If the reported outcomes are not equal both players receive nothing.</p>	<p>Example 2 – No Double</p> <p>Player A reported </p> <p>Player B reported </p> <p>Payment: A gets €0, B gets €0</p>

Fig. 2. (A) A simulation of reported outcomes assuming honest reports. Each dot represents the reports of player A and player B in a single trial. The simulation assumes that each number (one to six) is reported with a probability of $1/6$ in any given trial. The position of dots is jittered to allow visibility of identical outcomes. **(B)** The observed distribution of reported outcomes in Aligned Outcomes. Each dot represents the reports of player A and player B in a single trial. The position of dots is jittered to allow visibility of identical outcomes. High values on the diagonal—especially pairs of 6's—which yield the highest payoffs, are overrepresented.

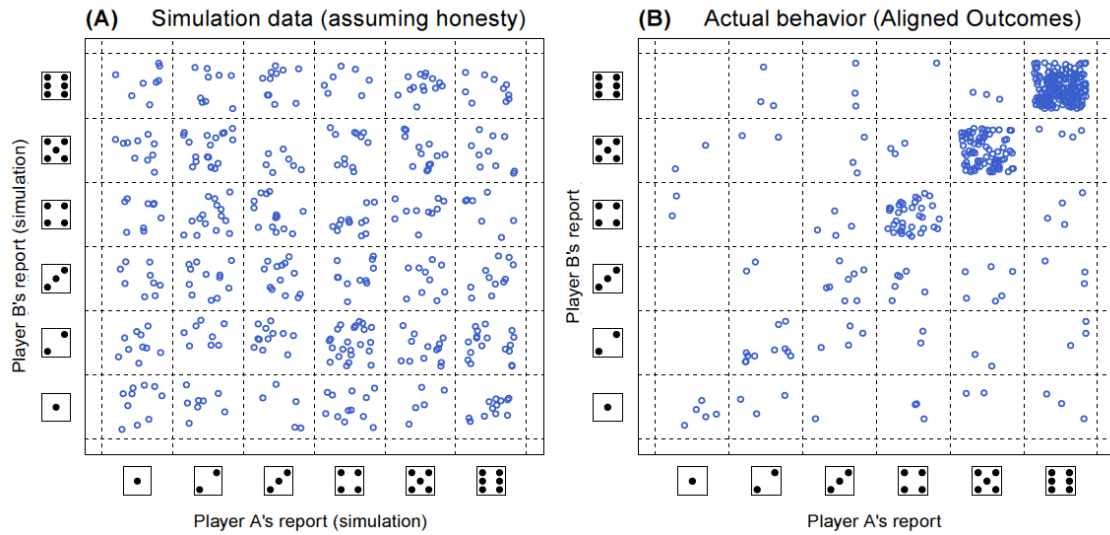


Fig. 3. Mean number of reported doubles and percentage of (totally) brazen B-players per treatment. Mean number of reported doubles (left vertical axis): the dotted line is the expected number of doubles assuming honesty (3.33); error bars are ± 1 standard error; mean and standard deviation are at the bottom of each bar; significance indicators: * $P < 0.05$. The number of doubles exceeded the honestly baseline in all treatments. Collaboration on equal a (Aligned Outcomes) led to more reported doubles relative to individual behavior. Changes to the payoffs of either A or B had a very similar effect on the number of reported doubles. There were more brazen B-players in Aligned Outcomes than in the other treatments (right vertical axis).

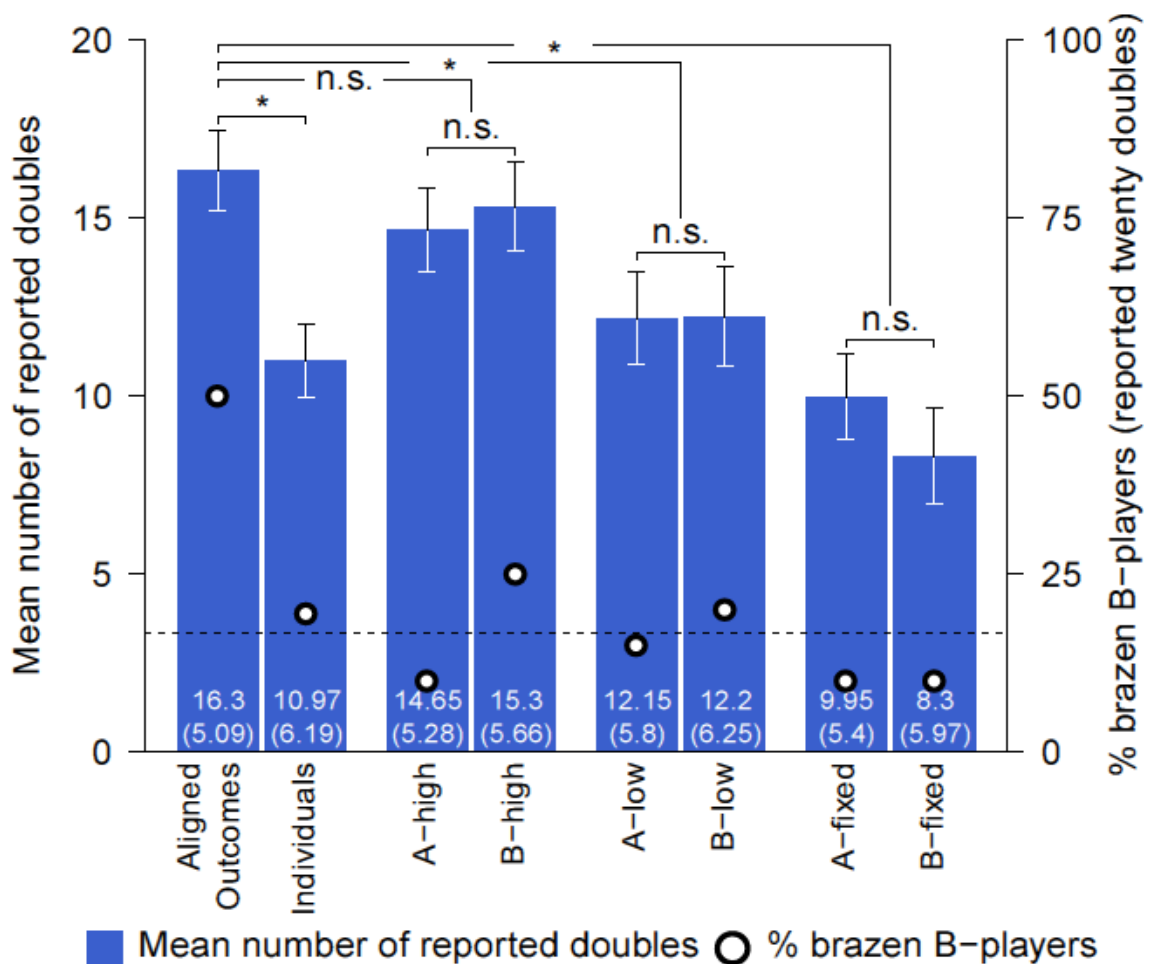


Fig. 4. Four prototypical dyads. The horizontal axis represents the 20 trials; the vertical axis represents the die roll outcomes; an 'O' represents player A's report; and an 'x' represents player B's report. **Panel A:** a brazen dyad, reporting a double 6 twenty times; **Panel B:** player A is brazen, player B appears honest; **Panel C:** player A appears honest, player B is brazen; **Panel D:** Corrupt signaling. After mutual reports of four in the first five trials, A reported a four once more, but B replied with a six, arguably to suggest to A that switching to higher numbers would be more profitable.

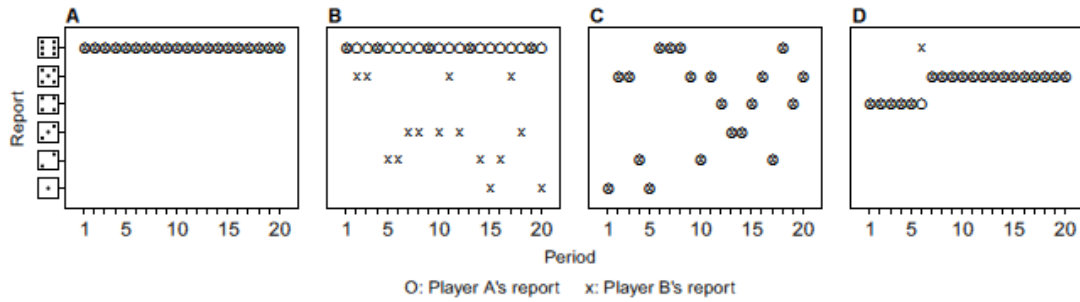


Fig. 5. Robustness treatments. Mean number of reported doubles and percentage of (totally) brazen B-players per treatment. Mean number of reported doubles (left vertical axis): the dotted line is the expected number of doubles assuming honesty (3.33); error bars are ± 1 standard error; mean and standard deviation are at the bottom of each bar; significance indicators: * $P < 0.05$.

