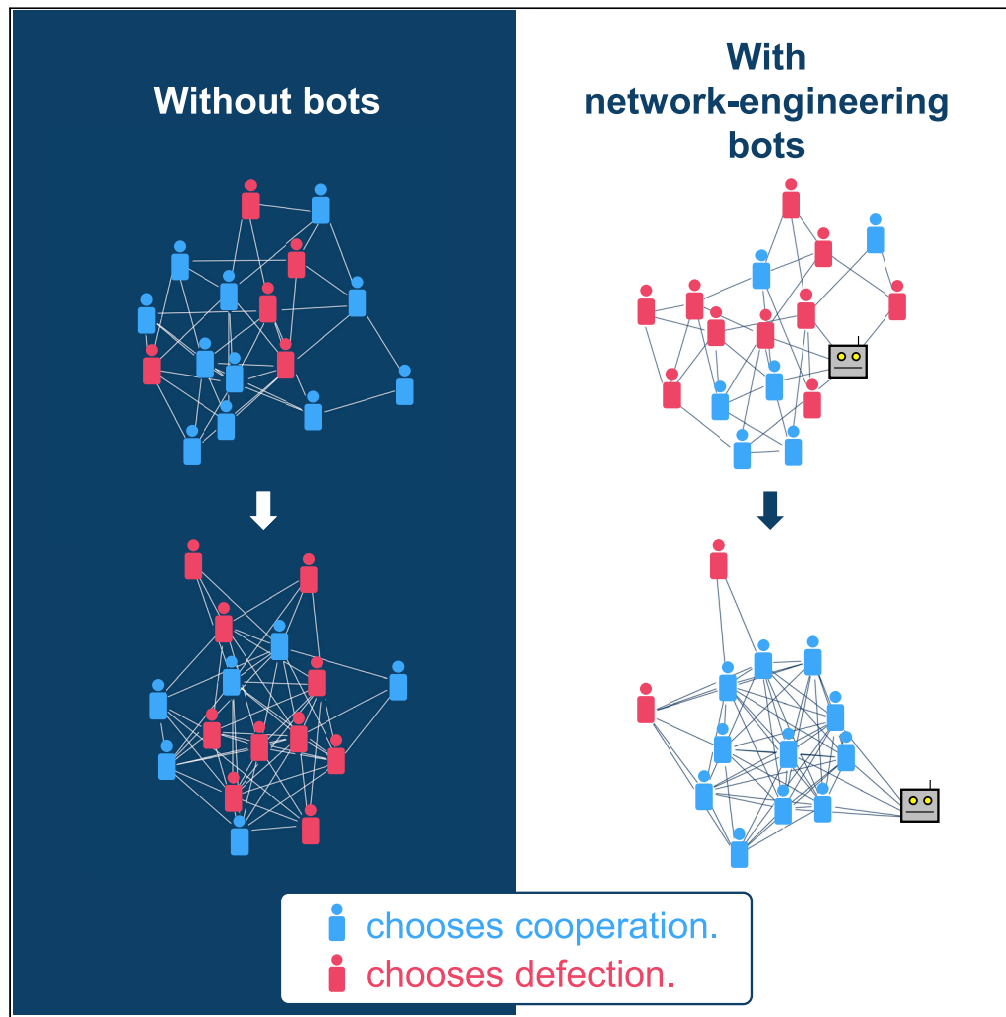


Article

# Network Engineering Using Autonomous Agents Increases Cooperation in Human Groups



Hirokazu Shirado,  
Nicholas A.  
Christakis

shirado@cmu.edu

**HIGHLIGHTS**

Network experiments show that adding bots can promote cooperation in human groups

To promote cooperation, bots intervene locally in the connections between humans

Even a single bot, with simple AI, can foster cooperation by network engineering

Network-engineering bots are effective even when identified as bots

Shirado & Christakis, iScience  
23, 101438  
September 25, 2020 © 2020  
The Authors.  
<https://doi.org/10.1016/j.isci.2020.101438>



## Article

## Network Engineering Using Autonomous Agents Increases Cooperation in Human Groups

Hirokazu Shirado<sup>1,6,\*</sup> and Nicholas A. Christakis<sup>2,3,4,5</sup>

## SUMMARY

Cooperation in human groups is challenging, and various mechanisms are required to sustain it, although it nevertheless usually decays over time. Here, we perform theoretically informed experiments involving networks of humans (1,024 subjects in 64 networks) playing a public-goods game to which we sometimes added autonomous agents (bots) programmed to use only local knowledge. We show that cooperation can not only be stabilized, but even promoted, when the bots intervene in the partner selections made by the humans, reshaping social connections locally within a larger group. Cooperation rates increased from 60.4% at baseline to 79.4% at the end. This network-intervention strategy outperformed other strategies, such as adding bots playing tit-for-tat. We also confirm that even a single bot can foster cooperation in human groups by using a mixed strategy designed to support the development of cooperative clusters. Simple artificial intelligence can increase the cooperation of groups.

## INTRODUCTION

Human societies function best when people produce public goods that offer collective benefits that they could not otherwise obtain individually (Olson, 1965). The cooperation required to do this, however, is challenging because it creates a social dilemma (Axelrod, 1984; Dawes, 1980): the group does well if individuals cooperate, but, for each individual, there is a temptation to defect. Getting groups to cooperate, and to keep cooperating, therefore presents substantial difficulties (Hardin, 1968; Nowak, 2006).

Various strategies for individuals to minimize their own losses or to act fairly, from their own point of view, when facing a cooperation dilemma have been identified. And these actions might conceivably also sustain cooperation in groups as a whole. For instance, in a repeated game, individual strategies—such as Tit-For-Tat (Axelrod, 1984) and its variants (Hilbe et al., 2013; Nowak and Sigmund, 1993)—may lead another person to whom a person is connected to cooperate as a secondary effect. In fact, even robots can be programmed to elicit cooperation from humans (Crandall et al., 2018). Furthermore, a large body of work has explored broader, institutional approaches to overcoming cooperation dilemmas, such as reputation (Cuesta et al., 2015; Nowak and Sigmund, 2005), punishment (Fehr and Gächter, 2002; Fowler, 2005), rewards (Rand et al., 2009), population structure (Allen et al., 2017; Ohtsuki et al., 2006), tie rewiring (Rand et al., 2011), or the establishment of a central authority (Ostrom, 1990). But it is still unclear how a small fraction of individuals might guide a group of others toward the creation of public goods *without* a super-ordinate institutional change. And approaches that actually increase levels of cooperation in groups (from their baseline) are scant.

Here, we examine how individual autonomous agents acting locally can facilitate, and even increase, cooperation in a group of people. We focus on network interventions (Valente, 2012) on the assumption that all individuals, including those who attempt to intervene to make the situation better, are embedded in a social network and that their information and actions are limited to their local neighborhood of connections (Granovetter, 1985). We take preprogrammed autonomous agents endowed with various simple strategies (i.e., “bots” or computer programs behaving as individual actors in a social system) and introduce them into the network and have them interact as members of the group (Shirado and Christakis, 2017). We do not assume that the bots have distinctive advantages that allow them to observe and change the entire social system as a central authority might. As a result, we can (1) study the abstract principles instantiated by the bots (gaining insights into how similar actions initiated by humans might affect cooperation in groups, albeit with exquisite experimental control in the case of the bots) and (2) develop practical applications

<sup>1</sup>School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

<sup>2</sup>Yale Institute for Network Science, Yale University, New Haven, CT 06520, USA

<sup>3</sup>Department of Sociology, Yale University, New Haven, CT 06520, USA

<sup>4</sup>Department of Ecology & Evolutionary Biology, Yale University, New Haven, CT 06511, USA

<sup>5</sup>Department of Biomedical Engineering, Yale University, New Haven, CT 06520, USA

<sup>6</sup>Lead Contact

\*Correspondence: shirado@cmu.edu

<https://doi.org/10.1016/j.isci.2020.101438>



of distributed, simple artificial intelligence (AI) agents in the form of bots that might be introduced into online groups so as to modify their properties for the better (Paiva et al., 2018). In short, we explore how bots can be used to facilitate positive social outcomes.

Humans decide whether to cooperate in part based on the actions of their neighbors (Shirado et al., 2013). Thus, an intervening agent, which resembles what Axelrod has called a “reformer” (Axelrod, 1984) (here, one that it is embedded within the system as a participant itself), might, by exercising the behavioral options of cooperating or defection, not only affect its own payoffs but also have an impact on others. However, a reformer might be limited in its ability to make its own network environment favorable to cooperation (Liu et al., 2011). For instance, theory suggests that the simple strategy of continuous cooperation by a well-intentioned reformer may do little to change cooperation dynamics, especially in large groups (Olson, 1965), and any such cooperators might be solely exploited by the defectors connected to them (Axelrod, 1984).

To address the challenges faced by reformers embedded in social networks who seek to sustain or enhance group cooperation, we focus on network dynamics that allow individuals to adjust social ties and engage in a kind of social network engineering (Rand et al., 2011; Shirado et al., 2013). That is, we explore a strategy whereby each bot gives its human neighbor an option to make or break a tie with other human subjects chosen by the bot (the bots act as a kind of social assistant). To find an effective strategy for such reformer bots, we set up our experiments in two stages. In the first set of experiments, we examine various network interventions (including a variety of alternative, control strategies) using a sufficient number of bots (“Experiment 1”). We evaluate the efficacy of the intervention in a repeated set of interactions in a public goods game involving cooperation. In the second experiment, we specify the bot’s intervention strategy based on the results of the first experiment and then test it with a single bot embedded in a group of human subjects (“Experiment 2”).

In Experiment 1, we recruited 896 unique human subjects through Amazon Mechanical Turk, dividing them among 56 groups, in sessions lasting an average of 24.1 min. Subjects were placed into groups of 16 individuals arranged in a network with an Erdős-Rényi random graph configuration (Erdős and Rényi, 1959) in which 30% of the possible ties were present at the outset, on average. In the sessions involving bots, each subject additionally received one connection with a bot. The subjects were therefore initially connected to an average 4.41 (SD = 1.76) other humans and 1 bot (i.e., 5.41 network neighbors on average). Subjects could identify each neighbor by a permanent, randomly generated name (see [Data S1](#) and [Transparent Methods](#)).

Each subject played a public-goods game lasting 30 rounds with their network neighbors without knowing when the game would end. At the beginning of the game, subjects received US\$1.00 as their initial endowment. In each round, all the subjects chose whether to cooperate, by reducing their own endowment US\$0.05 per neighbor in order to increase the endowment of all their neighbors by US\$0.10 each, or to defect, by paying no cost and providing no benefits. Subjects had to make the same choice with respect to all their connected neighbors (Allen et al., 2017; Cuesta et al., 2015; Ohtsuki et al., 2006; Rand et al., 2011, 2014; Shirado et al., 2013; Yamagishi, 1986).

After making their cooperation choice, subjects were informed of the choices made by their neighbors. Then, subjects had the opportunity to change their neighbors by making or breaking ties (“tie-rewiring” options), using a game setup that we had explored in prior experiments (Rand et al., 2011; Shirado et al., 2013). Specifically, 5% of all pairs of human subjects (i.e., 6 pairs, on average) were chosen at random in each round and were given the opportunity to rewire their ties. If a tie already existed between the two subjects, then one of the two was picked at random to be allowed to choose whether to voluntarily break the tie with the other; if a tie did not already exist between the two, a randomly selected member of the pair was given the option to form the tie (unilaterally). When making this decision, subjects were aware of whether the person to whom they might disconnect or connect had cooperated or defected in the past round. Thus, people could choose to modify a subset of their social ties (with low frequency) at each round; the network could become rewired as a result of these modifications; and all subjects’ network properties (such as network degree and the fraction of cooperators among their neighbors) could change over time. Our focus was on the possible impact of network interventions with bots placed within these groups.

Within this basic setup, we introduced 16 bots into the network of 16 human subjects (except for the control sessions without any bots). Each bot had only one tie and was connected with a different subject at the beginning of the game (i.e., every subject initially had precisely one tie to a bot among their neighbors). Bots always chose cooperation in the game (except for the bots using the Tit-For-Tat strategy). And the bots never connected with each other.

To be clear, we used the artifice of single-tie bots for Experiment 1 so as to experimentally fix the total amount of intervention across sessions and treatments to 16 total ties with bots at all times and in all treatments. Sixteen single-tie bots corresponded to three human subjects in terms of average initial network degree, and the human-bot connections accounted for 11.8% of all the possible connections in a network (i.e., 16/136). Subjects were not informed that there were bots in the game (except for an extra condition making bots visible; see below).

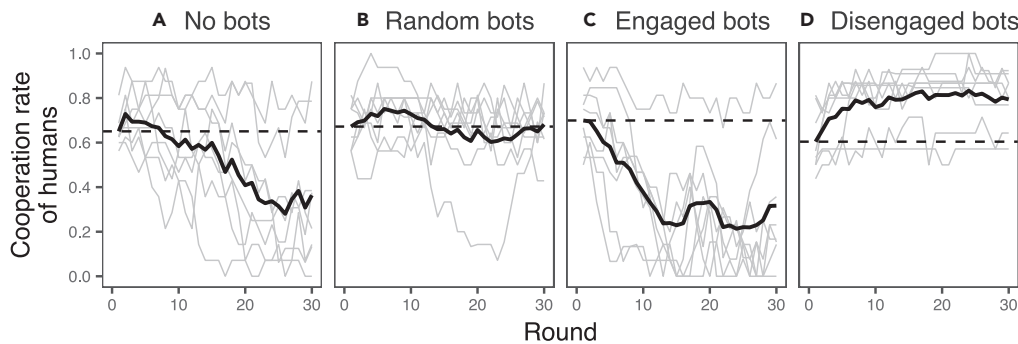
Using the bots embedded in a network of human subjects, we tested the network-intervention strategy whereby each bot generated an additional rewiring option for its human neighbor (regardless of whether the neighbor cooperated or defected) to make or break a tie with other human subjects; the humans did not have to use the rewiring option given by the network-intervention bots. All the network-intervention strategies added the same amount of rewiring options to the social system (increasing the possible rewiring rate from 5% to 16%).

We then manipulated the criteria used by the bot reformers to pick targets. The bots intervened based on the cooperation decision that the humans had previously made. To improve the level of cooperation in a group using network engineering, an intervention would need not only to retain existing cooperators, but also to prompt defectors to change their decisions. Thus, we tested three processes, in terms of whether the approach is *random*, *engaged*, or *disengaged*—from the reformer’s perspective—with respect to the defectors in the group (Rand et al., 2009). In the “random” process, targets are chosen at random. The target subject thus had an additional chance to form or break a tie with another subject, irrespective of the target subject’s cooperation decision. In the “engaged” process, defectors are given a chance to form more cooperative ties so as to (hopefully) encourage them to switch to cooperation. The engaged process had every bot give an option to its neighbor to make a new tie with another subject who was choosing cooperation (so, for example, a defector connected to the bot was introduced to a cooperator). That is, the cooperative bots foster engagement with the defectors (i.e., the bots are conciliatory to defectors). In the “disengaged” process, the cooperative bots foster disengagement with defectors (i.e., the bots work to quarantine defectors from cooperators). The disengaged process involved the bot giving an option to its neighbor to cut an existing tie with another subject who was choosing defection.

We also conducted sessions with three types of control conditions. The first control condition did not involve any bots. In the other two control conditions, like the treatment conditions, the group of 16 human subjects had 16 single-tie bots in a network. In one control condition, the bots (again) always cooperated. In the other, they used the Tit-For-Tat strategy; they started with cooperation and then selected the same behavioral choice (i.e., cooperation or defection) as their human neighbor had selected in the previous round (Axelrod, 1984). Importantly, and in contrast to the treatments, the bots in these control conditions did not intervene in the local networks of the human subjects by giving them a rewiring option.

As noted, subjects were not informed that they were interacting with bots and that some of the tie-rewiring options were given by bots. To assess the effect of this ignorance, we also carried out experiments with an extra condition of “bot visible.” In this extra condition, subjects were informed that they were interacting with bots, which neighbors were bots, and which tie rewiring options the bots suggested them (see [Data S1](#) and [Transparent Methods](#)). Thus, subjects could make a decision with knowledge that they had both bots and humans among their neighbors and that the bots gave them rewiring options. We examined this treatment only in the case in which the bots used the “disengaged” strategy.

In Experiment 2, which involved an additional 128 subjects in 8 sessions, we tested the minimal intervention of a single bot. This experimental setting differed from Experiment 1 only in that a group of 16 human subjects played a public-goods game with 1 bot having 5 ties, which corresponded to 1 average human subject in terms of initial network degree (in the sessions of Experiment 2, the average network degree = 5.13 at round 1). Thus, in contrast of Experiment 1, some subjects had a connection with the bot and the others



**Figure 1. The Fraction of Cooperative Human Players per Round**

Light gray lines show results for each session, black lines show average across all experimental sessions for each treatment ( $N_{\text{session}} = 8$  per treatment). Initial rates of average cooperation varied by chance across treatments (the dashed lines). See the results of two other control conditions (“always cooperate” and “tit-for-tat”) in Figure S1. Across all 48 groups, the average initial rate of cooperation was  $68.2\% \pm 12.8\%$ .

did not (in every round). This single bot used a network-engineering strategy that we designed according to the results of Experiment 1 (as described below).

In sum, we evaluated the effect of bot interventions and network engineering with two experiments. Experiment 1 had 3 control conditions not involving any network-intervention bots, 3 treatment conditions (i.e., random, engaged, and disengaged criteria for how the bots treated defectors), and 1 extra condition involving manipulation of bot visibility. Experiment 2 had one condition involving a single bot. We conducted 8 sessions for each condition for a total of 64 groups with 1,024 human subjects in total.

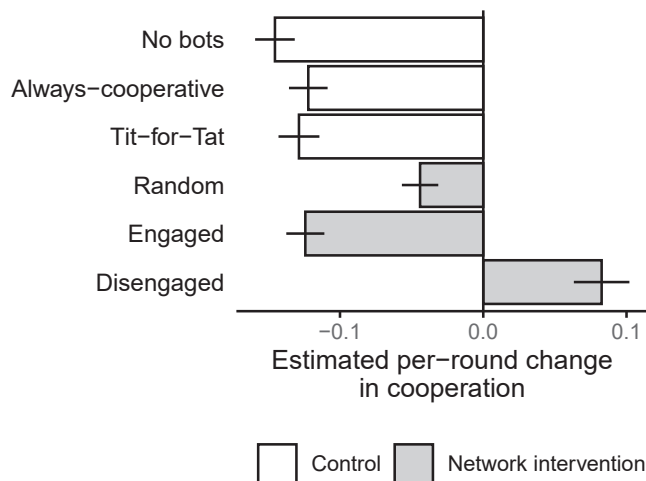
Subjects interacted anonymously over the Internet using our publicly available software platform (available at [breadboard.yale.edu](http://breadboard.yale.edu)). We allowed the participation only of those subjects who completed a tutorial session and passed a series of tests assessing their understanding of the game rules. We prohibited subjects from participating in more than one session. All subjects consented, and this study was approved by the Yale University Committee of the Use of Human Subjects.

## RESULTS

### Improving Human Cooperation with Disengaged Intervention Bots (Experiment 1)

For the control sessions involving only human subjects ( $N_{\text{subject}} = 16$  per session), on average, 65.1% of subjects per session started with choosing cooperation, and the fraction of cooperative players decreased over the rounds, eventually reaching 36.4% at the final round ( $p = 0.024$ ; paired t test with  $N_{\text{session}} = 8$  sessions; Figure 1A). This finding is in keeping with much prior work showing that defection overwhelms cooperation in human groups as interactions progress over time (Axelrod, 1984; Dawes, 1980; Nowak, 2006). These social dynamics favoring defection did not change materially when either the always-cooperative or Tit-For-Tat (TFT) bots were added to the group (Figure S1A). These findings illustrate the limited ability of such reformers’ actions to help groups to maintain cooperation, whatever the relative value of these strategies for the actors themselves who play them. That is, TFT bots, for instance, did not help the groups in which they were embedded.

When bots using the network interventions were introduced to the group, the declining slope in cooperation depended on the criterion for rewiring that the bots used. When bots randomly gave tie-rewiring options to humans in their group (random intervention), the cooperation level of human subjects stayed steady (Figure 1B; 67.2% at the outset and 68.1% at the end, on average;  $p = 0.872$ ; paired t test with  $N_{\text{session}} = 8$ ). When bots gave the option for the humans to make a tie only to other subjects who chose cooperation (engaged intervention), cooperation decayed quickly from 69.9% and stabilized at 31.7% (Figure 1C;  $p < 0.001$ ; paired t test with  $N_{\text{session}} = 8$ ). In contrast, when bots gave the option to the humans to break a tie to other subjects choosing defection (disengaged intervention), cooperation improved from the initial status and stabilized at a high level of cooperation (Figure 1D; 60.4% at the outset and 79.4% at the end, on average;  $p = 0.020$ ; paired t test with  $N_{\text{session}} = 8$ ). We are not familiar with any prior work documenting interventions that have been able not just to



**Figure 2. Average Change in Rates of Cooperation by Round**

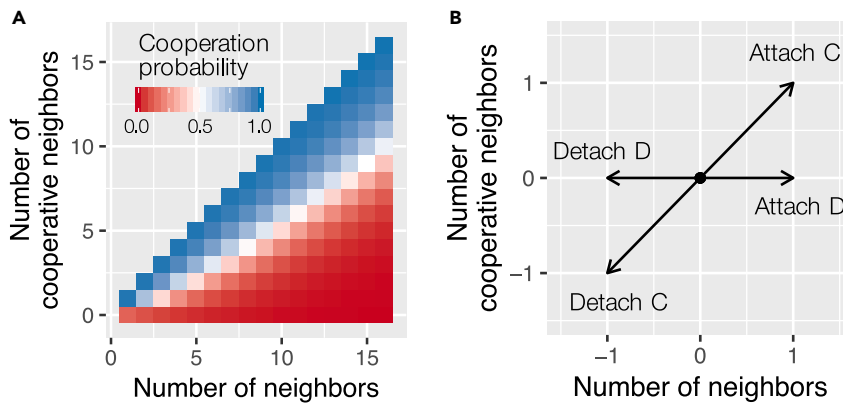
Estimates based on GLMM, using a logistic regression model of individual cooperation choice with random effects for session and individuals (see [Transparent Methods](#)). The error bars are 95% confidence interval (CI).

achieve the maintenance of cooperation in groups involved in public goods scenarios, but also to effectuate an increase from the baseline rate of cooperation within a group, let alone using an approach involving network engineering with autonomous agents.

We comprehensively evaluated how cooperation improved or decayed over the rounds across our six conditions. We conducted this analysis at the level of individual cooperation decisions using a generalized linear mixed model (GLMM) with random effects for sessions with nested individuals (see [Transparent Methods](#)). [Figure 2](#) shows the relationship between the bot intervention strategy and individual-level cooperation. In regression models, only the network-intervention treatment using the disengaged (defector-detachment) criterion had a positive effect on the cooperation probability of individuals ( $p < 0.001$ ; GLMM with  $N_{\text{subject}} = 768$  in  $N_{\text{session}} = 48$ ; 6 conditions  $\times$  8 sessions  $\times$  16 subjects). The other intervention strategies of the bots, such as TFT, failed to change the course of decision-making that otherwise favored defection in individuals across time ([Figure 2](#)). Moreover, a comparison of the disengaged strategy with each of the other strategies shows a significant improvement over those alternatives ( $p < 0.001$ ; GLMM with  $N_{\text{subject}} = 768$  in  $N_{\text{session}} = 48$ ; see [Table S1](#)).

The overall average number of ties per human generally rose from the start in all conditions except for the sessions with disengaged intervention bots ([Figure S1B](#)). The disengaged intervention bots initially reduced the connections by offering tie-rewiring options to break connections from non-cooperative neighbors. Subjects did increase their connections and develop cooperative behavior after a couple of rounds, and the density reached the same level as that of the sessions without bots by the last round ( $p = 0.350$ ;  $t$  test with  $N_{\text{session}} = 16$ ); this also had implications for total contributions ([Figure S1C](#)).

The disengaged intervention bots that fostered detachment from defectors helped subjects modify their neighborhood environment in ways that made it easier for the subjects to choose cooperation. We analyzed the network effects on cooperation decisions in human subjects using GLMM with random effects for individuals (see [Transparent Methods](#) for details). [Figure 3A](#) shows the estimated cooperation probability of a participant depending on the number of total neighbors and the number of cooperative neighbors (the interaction term also captures the *fraction* of cooperative neighbors). The individual-level analysis shows that network degree (i.e., the number of local neighbors) negatively affects the probability of a subject choosing cooperation, whereas the number and rate of cooperative neighbors have a positive impact on the cooperation probability ([Table S2](#)). In the setting of network interventions, adding or removing ties with defectors can cause a significant shift in individuals' cooperation probability ([Figure 3B](#)). Thus, the *engaged* intervention bots can discourage cooperators by helping defectors to attach to them ([Figure 1C](#)). On the other hand, the *disengaged* bots can encourage subjects (including even those who have chosen defection) to choose cooperation by helping them to detach from defectors ([Figure 1D](#)).



**Figure 3. Cooperation Pattern with Neighborhood Change**

(A) Cell color shows cooperation probabilities estimated by GLMM shown in Table S2.

(B) The impact of possible changes in a subject's surroundings is shown schematically. From a particular point in the parameter space (as shown in A), a person could move sideways or along the diagonal in the probabilistic space of human cooperation. Detaching or attaching to a defector changes the number of neighbors but does not change the number of cooperative neighbors; this direction obliquely crosses the contour of the cooperation probability distribution. On the other hand, detaching or attaching to a cooperator changes both the number of neighbors and the number of cooperative neighbors; such a change runs almost parallel to the contour of the cooperation probability distribution.

Furthermore, in our experimental conditions, only the disengaged intervention bots helped cooperators to cluster (Figure S1D). As the homophilic clusters of cooperators developed in the disengaged intervention condition, cooperators gradually outperformed defectors in terms of earnings (Figure S2). In the other treatments, the network dynamics were not enough to make cooperators outperform defectors (in terms of individual economic performance).

Using a further experiment, we confirmed that the effect of the disengaged intervention is obtained even when subjects knew whether they were interacting with bots, even when subjects knew which particular neighbors were bots, and even when subjects knew which tie rewiring options the bots had suggested to them (Figure S3). Subjects' awareness of bots and their interventions did not undermine the efficacy of the bots in this experimental setting.

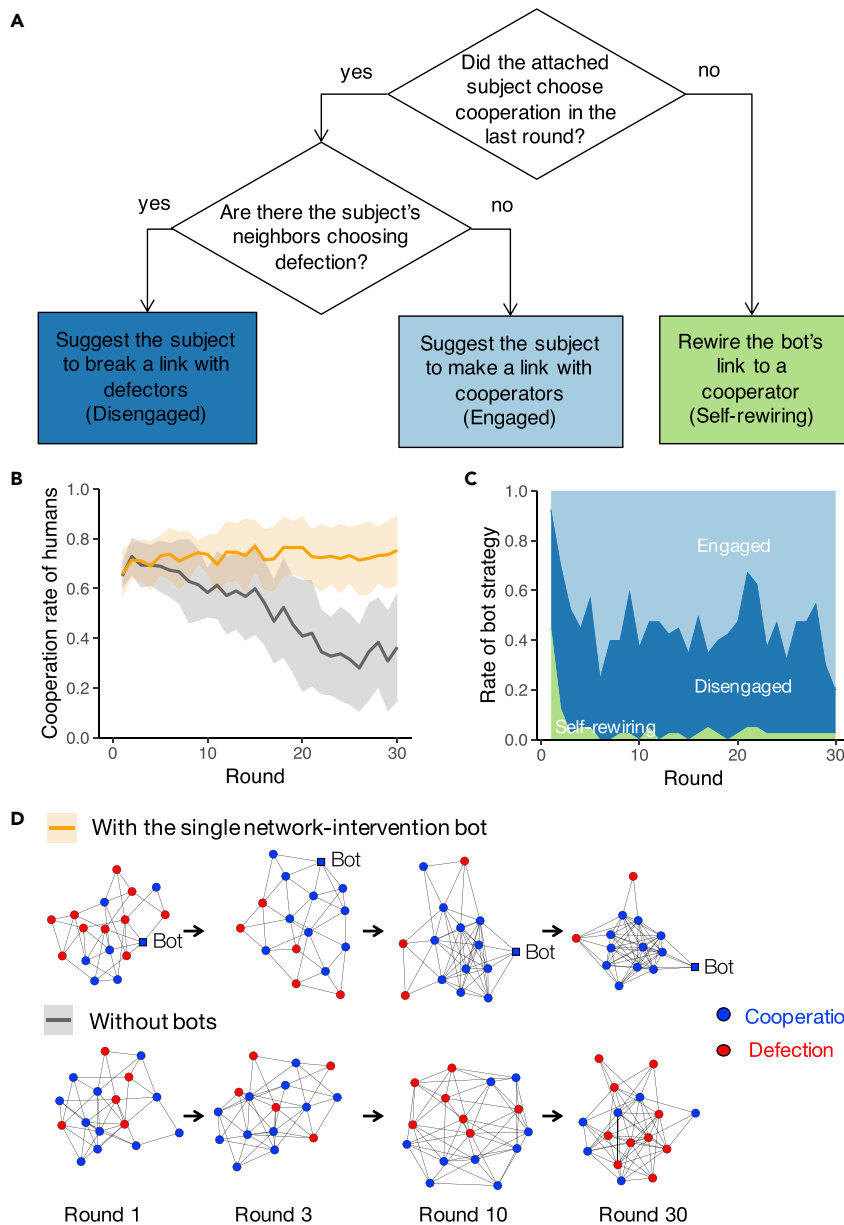
### Substantial Impact of A Single Bot (Experiment 2)

The results of Experiment 1 suggest that successful interventions involving disengaged intervention bots may support the development of cooperative clusters. Thus, we tested a network-intervention strategy that was designed to support subjects to build cooperative clusters. In Experiment 2, we introduced 1 bot having 5 ties to a network of 16 human subjects (where the average degree was 5.13 at round 1).

This single bot deployed a slightly more complicated, mixed strategy (Figure 4A). If the bot's subjects (i.e., the humans to which it was directly connected) had chosen defection in the last round, the bot would detach and connect to other subjects who were cooperators in the last round (chosen at random). But if the bot's subjects had chosen cooperation in the last round, what the bot did next would depend on what the subjects' neighbors had previously done. If the subjects had a neighbor who had chosen defection, the bot suggested that the subject break their link to such a neighbor (or offered them the chance to do so), i.e., the disengaged strategy. (If the subjects had more than one neighbor who had chosen defection, the bot chose one of those neighbors at random to suggest that they be abandoned.) If, on the other hand, all of the subjects' neighbors were cooperators, then the bot offered the subjects the chance to connect to another cooperator (beyond the current neighbors the subjects had), i.e., the engaged strategy. (And this had the further result of increasing the number of connections the subjects had.)

As a result, this single bot substantially changed the social dynamics in these groups and improved cooperation in a network of 16 humans (Figure 4B;  $p < 0.001$  with GLMM;  $N_{\text{subject}} = 128$  in  $N_{\text{session}} = 8$ ) while also fostering cooperative clusters (Figure S4). The bot initially sought out human partners through





**Figure 4. Cooperation and Network Dynamics with a Single Bot Deploying a Mixed Strategy**

(A) Control diagram for how a single bot intervenes in a network of human subjects.

(B) Experiment results regarding average cooperation fraction with 95% CI ( $N_{\text{session}} = 8$  for each treatment). The orange line indicates the result of sessions with the single network-engineering bot. The dark gray line indicates the result of sessions without bots, which is identical to Figure 1A.

(C) Experiment results regarding average rate of the bot's intervention strategy actually applied to human players, over the rounds.

(D) Network snapshots of an example session having a single bot and a session without bots.

“self-rewiring” and then suggested to the humans to which it had connected that they rewire their ties (Figure 4C). Remarkably, the single bot eventually employed the engaged strategy more than 50% of the time in its working to improve cooperation, but the same engaged strategy considerably *diminished* cooperation when it was the sole strategy used from the outset of the interactions in Experiment 1 (Figure 1C). This indicates that it is the bot's adaptation of its intervention strategy to fit the humans' situation that is indeed efficiently driving the increase in cooperation. In parallel with the shift in strategies, the bot itself moved to



the periphery of the network in the later rounds (Figure 4D). Given the initial interventions by the single bot, the humans came to foster cooperative circumstances *by themselves*, over time.

## DISCUSSION

We find that embedded autonomous agents that are equipped with a simple kind of AI and that constitute a relatively small fraction of the connections in a group can act as network engineers and enhance collective cooperation. They achieve this by distancing individuals from non-cooperative members of their groups and by fostering clusters of cooperators within groups, both of which encourage people to become more cooperative. Although prior work has empirically shown the possibility of *maintaining* cooperation using diverse interventions, including that all human players be allowed to fully control their partner selections (Hauge et al., 2019; Rand et al., 2011; Shirado et al., 2013), our experiment shows a way of *promoting* cooperation, indeed with a decentralized, individually implemented strategy involving network engineering by autonomous agents. The cooperation level in the group actually rises from the initial condition when bots that undertake the controlled defector-excluding network intervention strategy are introduced (Figure 2). That is, with the disengaged bots, individuals who may have initially been on course to exploit others changed their mind, which rarely occurs in human-only groups (Peysakhovich et al., 2014; Shirado et al., 2013). Moreover, such tie brokerage in individuals strongly affects cooperation in the group even when people can see that it is bots, not humans, who are intervening in their connections (Figure S3).

Cutting ties to defectors in our experiments can be seen as a form of decentralized ostracism, here abetted by the bot network engineers. This approach requires no changes in the underlying interaction system or the imposition of a central authority with a view of the whole system. The distributed bots neither consolidate information nor enforce rules. It builds on experiments with dynamic networks, where participants make their own decisions about whom to exclude from the benefits of individual cooperation (Rand et al., 2011; Shirado et al., 2013). We find that network-wide cooperation can indeed be increased using autonomous agents, including even a single individual agent that evolves its intervention strategies and leverages human behaviors regarding partner selection (Figure 4).

Models of evolutionary dynamics suggest that evolution may favor cooperative types when partner choice is feasible (Eshel and Cavalli-Sforza, 1982; McNamara et al., 2008; Stanley et al., 1993), and evidence from Hadza foragers (and other groups) suggests that people are strongly influenced in their cooperation behavior by the norms of the groups to which they belong (Smith et al., 2018). By endowing actors with the tendency to manipulate the ties around themselves, bots (and people) may be able to affect the cooperativity of their groups, creating a convivial local environment for cooperative behavior. It is known that the ties between an ego's neighbors (i.e., network transitivity) play an essential role in collective action (Centola and Macy, 2007). And the clustering of cooperators, whether induced by intervention (Rand et al., 2014) or occurring naturally (Apicella et al., 2012), is known to facilitate cooperation. Hence, personal introductions might be an important mechanism in the evolutionary process favoring cooperation (Nowak, 2006) and might play a role in confronting collective action problems (Olson, 1965). It is possible that humans might have required third-party mediation in their tie formation in parallel with evolving the capacity for cooperation (Purzycki et al., 2016), and there is some evidence that interpersonal variation in network transitivity may be heritable (Fowler et al., 2009). By engineering the tie formation between humans, autonomous agents can therefore help humans develop favorable environments for cooperation.

Our work indicates that it is possible to use simple computer programs to have meaningful effects on collective behavior because, in the social situations that concern us, the bots are mixed in with (much smarter) humans on a level playing field in what we call "hybrid systems" (Paiva et al., 2018; Rahwan et al., 2019; Shirado and Christakis, 2017; Traeger et al., 2020). The bots can function as a kind of social catalyst, helping humans to help themselves. Artificial-intelligence agents that are being developed to enhance the social good might not need to be extremely sophisticated, nor might they need to replace human cognition. Rather, they need only supplement human interaction. Since humans can change their own code of behavior by perceiving other people's behaviors (Axelrod, 1984), the simplicity and transparency of decision-making in artificial agents might make it even more intelligible to humans, thereby eliciting effective and stable effects (Nass et al., 1994).

Our work sheds light on techniques that might maintain or even increase rates of cooperation in groups, and it also offers the prospect of practical applications in online networks, where bot technology based

on people recommendation might be useful (Guy, 2018). For instance, it might be possible to improve the behavior of groups engaged in collective tasks like editing online materials (Kittur et al., 2007) or to reduce online harassment in social media (Johnson et al., 2019; Munger, 2017). On the other hand, bot interventions that result in high levels of transitivity might increase group-think and echo chamber effects (Stewart et al., 2019).

Our study offers empirical evidence that adding a few connections with autonomous agents in the role of reformers, mixed into human groups, can promote cooperation via a kind of network engineering. Although the results of laboratory experiments do not translate directly into the real world, the evidence suggests that agents that strategically intervene in local tie formation might actually increase cooperation in social networks—a notoriously difficult thing to do. Simple forms of artificial intelligence offer the prospect of improving human social interactions within groups.

### Limitations of the Study

Our work involves subjects interacting online in a highly stylized way. Moreover, there are features potentially relevant to cooperation interventions that our experiments do not explore, for example: how individuals would behave if they could choose cooperation separately for each partner (Melamed et al., 2018); whether individuals might recognize partners' attributes other than their cooperativity (Nishi et al., 2015); how accurate the information the bots rely on must be (Waniek et al., 2018); or whether the payoff structure (Ohtsuki et al., 2006), group size (Nosenzo et al., 2013), network topology (Allen et al., 2017), or the fraction and precise position of bot connections matter (Liu et al., 2011; Masuda, 2012). Not only agents' strategy but also agents' appearance and expression could affect human cooperation decisions (de Melo et al., 2018; Ju and Leifer, 2008). Another promising topic is developing more sophisticated approaches for the bots to learn the interaction strategies of humans (e.g., using deep learning techniques) (Vinyals et al., 2019). This study suggests that bots could influence social dynamics in human groups not only through their behavioral responses (typified by Tit-for-Tat and its variants) (Crandall et al., 2018), but also through their interventions in the social connections between people (Guy, 2018). Here, we only used simple algorithm in each bot. Bots might improve human cooperation more efficiently when they take a more flexible approach to adapt their intervention strategy to both individual and group situations. These are all important directions for future work.

### Resource Availability

#### Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Hirokazu Shirado ([shirado@cmu.edu](mailto:shirado@cmu.edu)).

#### Material Availability

This study did not generate new unique materials.

#### Data and Code Availability

The data and code in this manuscript are available at Mendeley Data: [<https://doi.org/10.17632/t963ktp6ft.1>].

## METHODS

All methods can be found in the accompanying [Transparent Methods supplemental file](#).

## SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at <https://doi.org/10.1016/j.isci.2020.101438>.

## ACKNOWLEDGMENTS

We thank F. W. Crawford, E. Erikson, B. Fotouhi, F. Fu, A. V. Papachristos, F. P. Santos, M. J. Crockett, and D. Rand for comments. M. McKnight provided expert programming needed for the online experiments. Funding for this research was provided by the Robert Wood Johnson Foundation, the NOMIS Foundation, Tata Sons Private Limited, Tata Consultancy Services Limited, and Tata Chemicals Limited.

## AUTHOR CONTRIBUTIONS

H.S. and N.A.C. designed the project. H.S. collected the data and performed the statistical calculations. H.S. and N.A.C. analyzed the results. H.S. and N.A.C. wrote the manuscript. N.A.C. obtained funding.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 22, 2020

Revised: July 13, 2020

Accepted: August 3, 2020

Published: September 25, 2020

## REFERENCES

- Allen, B., Lippner, G., Chen, Y.-T., Fotouhi, B., Momeni, N., Yau, S.-T., and Nowak, M.A. (2017). Evolutionary dynamics on any population structure. *Nature* 544, 227–230.
- Apicella, C.L., Marlowe, F.W., Fowler, J.H., and Christakis, N.A. (2012). Social networks and cooperation in hunter-gatherers. *Nature* 481, 497–501.
- Axelrod, R. (1984). *The Evolution of Cooperation* (Basic Books).
- Centola, D., and Macy, M. (2007). Complex contagions and the weakness of long ties. *Am. J. Sociol.* 113, 702–734.
- Crandall, J.W., Oudah, M., Tennom, Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., Cebrian, M., Shariff, A., Goodrich, M.A., and Rahwan, I. (2018). Cooperating with machines. *Nat. Commun.* 9, 233.
- Cuesta, J.A., Gracia-Lázaro, C., Ferrer, A., Moreno, Y., and Sánchez, A. (2015). Reputation drives cooperative behaviour and network formation in human groups. *Sci. Rep.* 5, 7843–7846.
- Dawes, R.M. (1980). Social dilemmas. *Annu. Rev. Psychol.* 31, 169–193.
- de Melo, C.M., Khooshabeh, P., Amir, O., and Gratch, J. (2018). Shaping Cooperation between Humans and Agents with Emotion Expressions and Framing. In *Proceedings of the 17th International Conference on Autonomous Agents and Multi Agent Systems*, 2224–2226.
- Erdős, P., and Rényi, A. (1959). On random graphs. *Publ. Math.* 6, 290–297.
- Eshel, I., and Cavalli-Sforza, L.L. (1982). Assortment of encounters and evolution of cooperativeness. *Proc. Natl. Acad. Sci. U.S.A.* 79, 1331–1335.
- Fehr, E., and Gächter, S. (2002). Altruistic punishment in humans. *Nature* 425, 137–140.
- Fowler, J.H. (2005). Altruistic punishment and the origin of cooperation. *Proc. Natl. Acad. Sci. U.S.A.* 102, 7047–7049.
- Fowler, J.H., Dawes, C.T., and Christakis, N.A. (2009). Model of genetic variation in human social networks. *Proc. Natl. Acad. Sci. U.S.A.* 106, 1720–1724.
- Granovetter, M. (1985). Economic action and social structure: the problem of embeddedness. *Am. J. Sociol.* 91, 481–510.
- Guy, I. (2018). People recommendation on social media. In *Social Information Access, vol 10100*, P. Brusilovsky and D. He, eds. (Springer), pp. 570–623, *Lecture Notes in Computer Science*.
- Hardin, G. (1968). The tragedy of the commons. *Science* 162, 1243–1248.
- Hauge, K.E., Brekke, K.A., Nyborg, K., and Lind, J.T. (2019). Sustaining cooperation through self-sorting: the good, the bad, and the conditional. *Proc. Natl. Acad. Sci. U.S.A.* 116, 5299–5304.
- Hilbe, C., Nowak, M.A., and Sigmund, K. (2013). Evolution of extortion in iterated prisoner's dilemma games. *Proc. Natl. Acad. Sci. U.S.A.* 110, 6913–6918.
- Johnson, N.F., Leahy, R., Restrepo, N.J., Velasquez, N., Zheng, M., Manrique, P., Devkota, P., and Wuchty, S. (2019). Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature* 573, 261–265.
- Ju, W., and Leifer, L. (2008). The design of implicit interactions: making interactive systems less obnoxious. *Des. Issues* 24, 72–84.
- Kittur, A., Suh, B., Pendleton, B.A., and Chi, E.H. (2007). He says, she says: conflict and coordination in Wikipedia. *SIGCHI* 07, 453–462.
- Liu, Y.-Y., Slotine, J.-J., and Barabási, A.-Á. (2011). Controllability of complex networks. *Nature* 473, 167–173.
- Masuda, N. (2012). Evolution of cooperation driven by zealots. *Sci. Rep.* 2, 646.
- McNamara, J.M., Barta, Z., Fromhage, L., and Houston, A.I. (2008). The coevolution of choosiness and cooperation. *Nature* 451, 189–192.
- Melamed, D., Harrell, A., and Simpson, B. (2018). Cooperation, clustering, and assortative mixing in dynamic networks. *Proc. Natl. Acad. Sci. U.S.A.* 115, 951–956.
- Munger, K. (2017). Tweetment effects on the tweeted: experimentally reducing racist harassment. *Polit. Behav.* 39, 629–649.
- Nass, C., Steuer, J., and Tauber, E.R. (1994). Computers are social actors. *SIGCHI* 94, 72–78.
- Nishi, A., Shirado, H., Rand, D.G., and Christakis, N.A. (2015). Inequality and visibility of wealth in experimental social networks. *Nature* 526, 426–429.
- Nosenzo, D., Quercia, S., and Sefton, M. (2013). Cooperation in small groups: the effect of group size. *Exp. Econ.* 18, 4–14.
- Nowak, M., and Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature* 364, 56–58.
- Nowak, M.A. (2006). *Evolutionary Dynamics* (Harvard University Press).
- Nowak, M.A., and Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature* 437, 1291–1298.
- Ohtsuki, H., Hauert, C., Lieberman, E., and Nowak, M.A. (2006). A simple rule for the evolution of cooperation on graphs and social networks. *Nature* 441, 502–505.
- Olson, M. (1965). *The Logic of Collective Action: Public Goods and the Theory of Groups* (Harvard University Press).
- Ostrom, E. (1990). *Governing the Commons* (Cambridge University Press).
- Paiva, A., Santos, F.P., and Santos, F.C. (2018). Engineering pro-sociality with autonomous agents. *AAAI* 18, 7994–7999.
- Peysakhovich, A., Nowak, M.A., and Rand, D.G. (2014). Humans display a “cooperative phenotype” that is domain general and temporally stable. *Nat. Commun.* 16, 4939.
- Purzycki, B.G., Apicella, C., Atkinson, Q.D., Cohen, E., McNamara, R.A., Willard, A.K., Xygalatas, D., Norenzayan, A., and Henrich, J. (2016). Moralistic gods, supernatural punishment and the expansion of human sociality. *Nature* 530, 327–330.
- Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J.W., Christakis, N.A., Couzin, I.D., Jackson, M.O., et al. (2019). Machine behaviour. *Nature* 568, 477–486.
- Rand, D.G., Arbesman, S., and Christakis, N.A. (2011). Dynamic social networks promote cooperation in experiments with humans. *Proc. Natl. Acad. Sci. U.S.A.* 108, 19193–19198.

Rand, D.G., Dreber, A., Ellingsen, T., Fudenberg, D., and Nowak, M.A. (2009). Positive interactions promote public cooperation. *Science* 325, 1272–1275.

Rand, D.G., Nowak, M.A., Fowler, J.H., and Christakis, N.A. (2014). Static network structure can stabilize human cooperation. *Proc. Natl. Acad. Sci. U S A* 111, 17093–17098.

Shirado, H., and Christakis, N.A. (2017). Locally noisy autonomous agents improve global human coordination in network experiments. *Nature* 545, 370–374.

Shirado, H., Fu, F., Fowler, J.H., and Christakis, N.A. (2013). Quality versus quantity of social ties in experimental cooperative networks. *Nat. Commun.* 4, 2814.

Smith, K.M., Larroucau, T., Mabulla, I.A., and Apicella, C.L. (2018). Hunter-Gatherers maintain assortativity in cooperation despite high levels of residential change and mixing. *Curr. Biol.* 28, 3152–3157.e4.

Stanley, E., Ashlock, D., and Tesfatsion, L. (1993). Iterated Prisoner's Dilemma with Choice and Refusal of Partners. ISU Economic Report Series, 9. [https://lib.dr.iastate.edu/econ\\_las\\_economicreports/9](https://lib.dr.iastate.edu/econ_las_economicreports/9).

Stewart, A.J., Mosleh, M., Diakonova, M., Arechar, A.A., Rand, D.G., and Plotkin, J.B. (2019). Information gerrymandering and undemocratic decisions. *Nature* 573, 117–121.

Traeger, M.L., Strohkorb Sebo, S., Jung, M., Scassellati, B., and Christakis, N.A. (2020). Vulnerable robots positively shape human

conversational dynamics in a human-robot team. *Proc. Natl. Acad. Sci. U S A* 117, 6370–6375.

Valente, T.W. (2012). Network interventions. *Science* 337, 49–53.

Vinyals, O., Babuschkin, I., Czarnecki, W.M., Mathieu, M., Dudzik, A., Chung, J., Choi, D.H., Powell, R., Ewals, T., Georgiev, P., et al. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 350–354.

Waniek, M., Michalak, T.P., Wooldridge, M.J., and Rahwan, T. (2018). Hiding individuals and communities in a social network. *Nat. Hum. Behav.* 2, 139–147.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *J. Personal. Soc. Psychol.* 51, 110–116.

**iScience, Volume 23**

**Supplemental Information**

**Network Engineering Using Autonomous Agents  
Increases Cooperation in Human Groups**

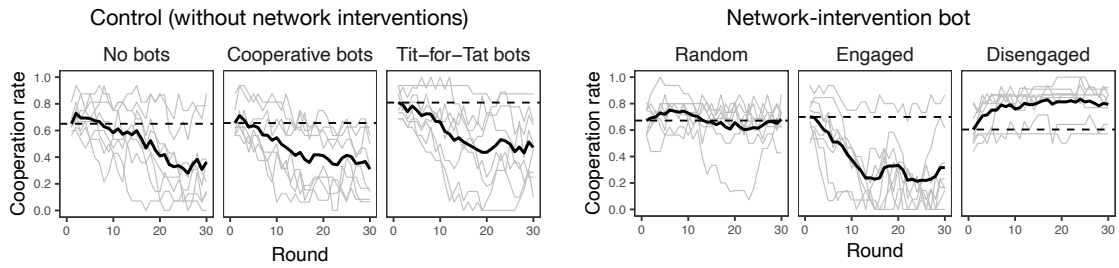
**Hirokazu Shirado and Nicholas A. Christakis**

## **Supplementary Information**

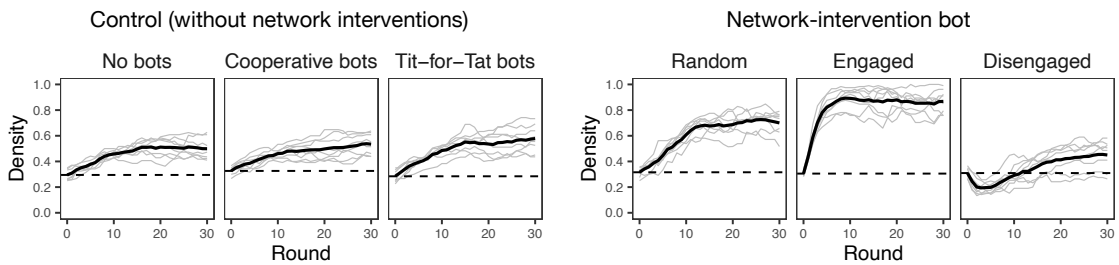
1. Supplementary materials
2. Transparent methods
3. Supplementary references

# 1. Supplementary materials

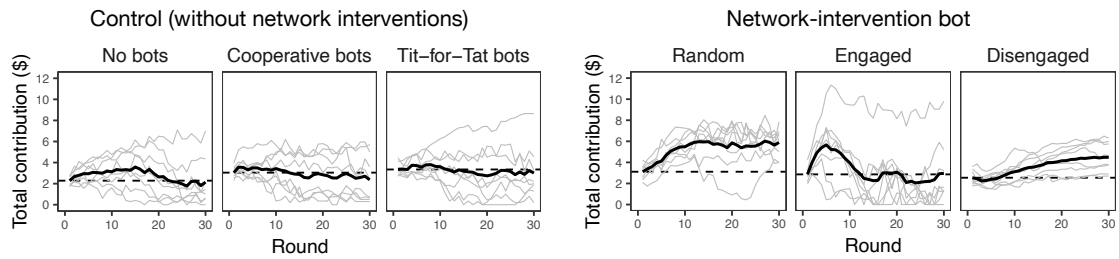
## A. Cooperation rate



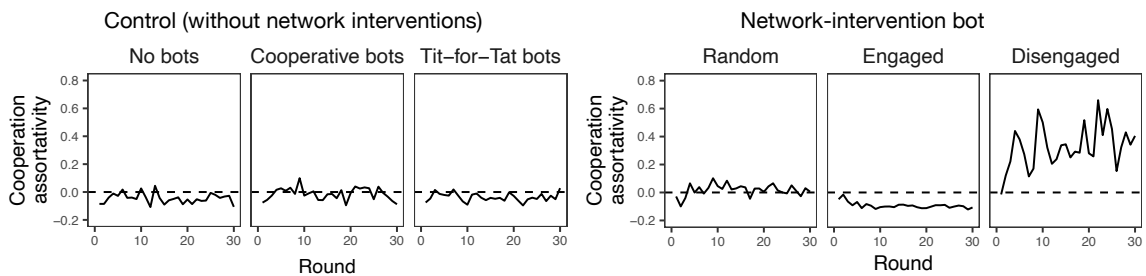
## B. Density



## C. Total contribution

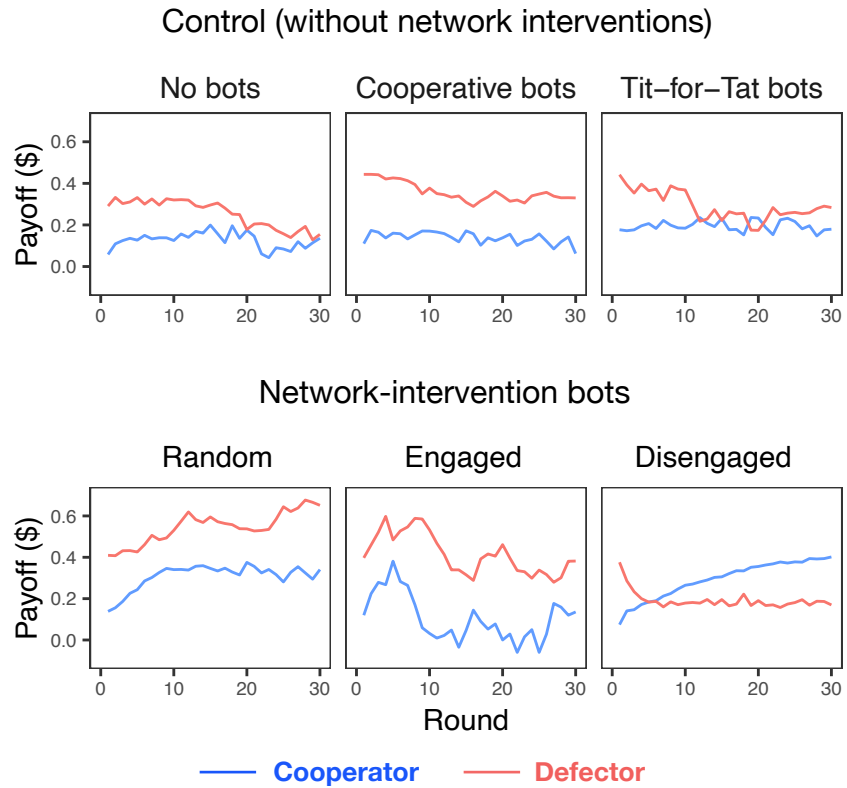


## D. Cooperation assortativity

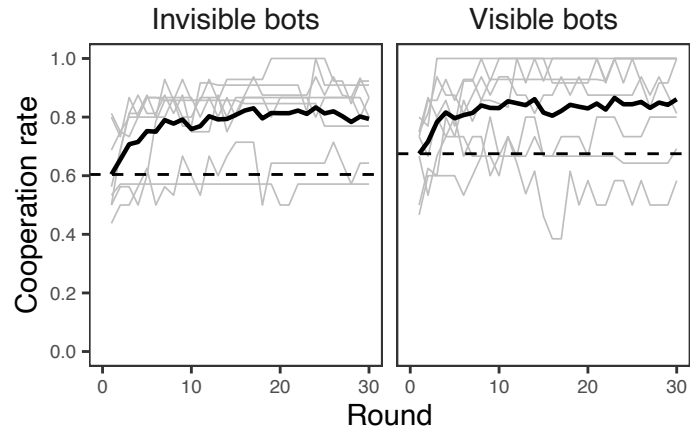


**Figure S1. Experiment results, Related to Figure 1.** (A) The fraction of cooperative human subjects by round. (B) Network density of human subjects by round. (C) Total amount of contribution from human subjects by round. Light gray lines show results for each session, black lines show average across all experimental sessions ( $N_{\text{session}}=8$  per treatment). Dashed lines show the initial average value per treatment. (D) Assortativity for cooperation in a human group by round. Cooperation assortativity is quantified by the correlation between pairs of connected nodes in terms of cooperation choice. Dashed lines show 0 coefficient of assortativity that indicates the same level of random connections regarding the behavioral choice.

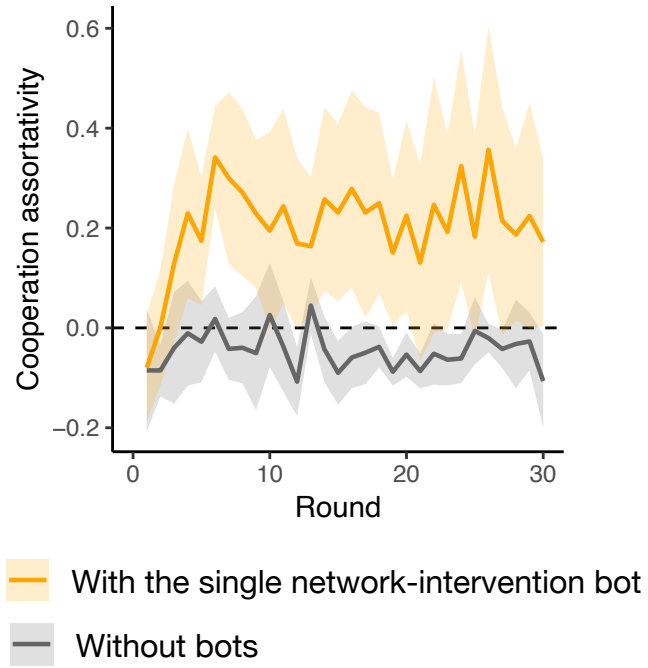




**Figure S2. Per-round payoff of human players by cooperation choice, Related to Figure 1.** Blue lines show the average payoffs that cooperating subjects received at the indicated round; red lines show those of defecting subjects.



**Figure S3. Fraction of cooperative human players by bot visibility, Related to Figure 1.** Light gray lines show results for each session, black lines show average across all experimental sessions for each treatment ( $N_{\text{session}}=8$  per treatment), and dashed lines show the initial rates of average cooperation. In the sessions with visible bots, humans were informed of which nodes were played by bots and which rewiring options were suggested by bots. In the session with invisible bots, humans were not informed (which is identical to Figure 1D). Bots intervened with the disengaged network-intervention strategy in both conditions.



**Figure S4. Change of cooperation assortativity in groups with a single bot using a mixed strategy, Related to Figure 4. The shades indicate 95% C.I. ( $N_{session}=8$ )**

**Table S1. The results of the statistical analysis regarding per-round cooperation change across bot treatments, estimated by GLMM with logit model incorporating random effects for sessions and individuals, Related to Figure 2.**

Intercept.....	2.122	***
	(0.300)	
Round.....	0.083	***
	(0.010)	
Round × Bot treatment (ref. Disengaged bots)		
No bots.....	-0.228	***
	(0.012)	
Always-cooperative bots.....	-0.205	***
	(0.012)	
Tit-for-Tat bots.....	-0.211	***
	(0.012)	
Random bots.....	-0.127	***
	(0.12)	
Engaged bots.....	-0.207	***
	(0.12)	
Number of observations.....	21146	

NOTE. Clustered standard errors are given in parentheses.

\*\*\*  $P < 0.01$ ; \*\*  $P < 0.05$ ; \*  $P < 0.1$

**Table S2. The results of the statistical analysis regarding cooperation probability, estimated by GLMM with logit model incorporating random effects for individuals, Related to Figure 3.**

	Model without status-quo effect	Model with status-quo effect
Intercept .....	1.221 *** (0.187)	0.797 *** (0.175)
Environement: Number of neighbors .....	-1.150 *** (0.216)	-1.073 *** (0.202)
Fraction of cooperators in neighbors .....	1.580 *** (0.258)	1.485 *** (0.241)
Number of cooperators in neighbors .....	1.216 *** (0.327)	1.065 *** (0.307)
Round .....	-0.393 *** (0.038)	-0.312 *** (0.038)
Self-action: From cooperation .....		0.593 ***
(ref. from defection)		(0.068)
Number of observations .....	20389	20389

NOTE1. The environment covariates and round number are standardized for estimation convergence.

NOTE2. Clustered standard errors are given in parentheses.

NOTE3. The data does not include the first round.

\*\*\*  $P < 0.01$ ; \*\*  $P < 0.05$ ; \*  $P < 0.1$

**Data S1. Instruction and tutorials, Related to Figure 1.** Below are screenshots for the initial description of the tutorial and the confirmation tests. We also show example screenshots of a real game.

The game will start in: 06:50

## Welcome!

You will be playing this game with other players. The task will start after a recruitment period. The game will begin when the time on the above Progress Bar elapses. You need to complete the tutorial and the comprehension test by then.

Please click 'Begin' to proceed to the tutorial. If you do not see a 'Begin' button below, please refresh your browser.

**Begin**



The game will start in: 06:41


## Tutorial 1/14

You are represented by the large circle. When you join the game, you will receive \$1.00. Your earnings will be indicated in your node to the left.

In the game, you will make decisions with other players that may cause you to gain or lose money.

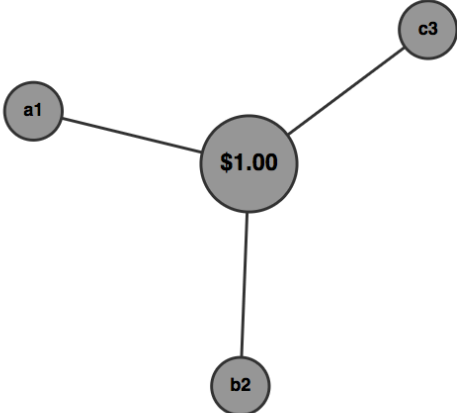
When you join the game and complete it, you will be paid the final earnings plus \$1.00 as completion bonus.

**Next**




## Data S1. Instruction and tutorials, Related to Figure 1. (cont.)

The game will start in: 06:33



**Tutorial 2/14**

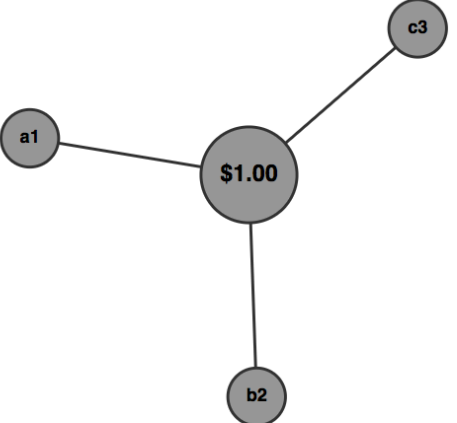
You and the other players will be arranged in a network. For example:



In this example, you have three partners directly connected to you in the network. **You will not see the whole network in the game. You will only see and interact with the partners you are directly connected to.**

**Next**

The game will start in: 06:25



**Tutorial 3/14**

You will be playing several rounds of this game; you will not be informed of the number of rounds in advance. Each round has two steps:

**Step 1. You choose whether to give money to your partners.**  
**Step 2. You choose to make or break connections with other players.**

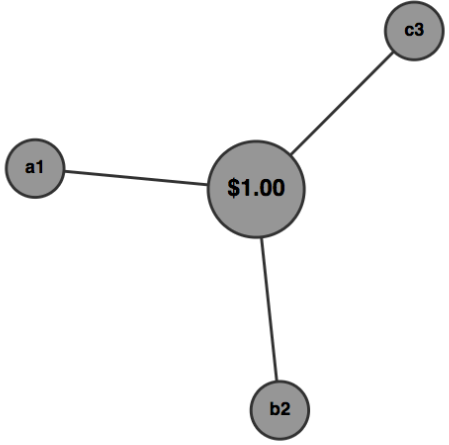
Other players will be making these same choices.  
We will now describe the game in more detail.

**Next**



## Data S1. Instruction and tutorials, Related to Figure 1. (cont.)

The game will start in: 06:16



**Tutorial 4/14**

You have two options in Step. 1:

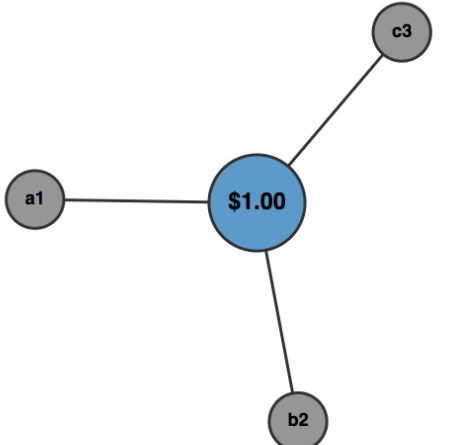
- If you **choose 'A'**, you keep your money for yourself.
- If you **choose 'B'**, you give \$0.05 to each partner; each of your partners earns \$0.10.

Other players have the same choice.

No matter what your partners choose, **you earn the most by keeping all of your money with option 'A'**.

**Next**

The game will start in: 06:07



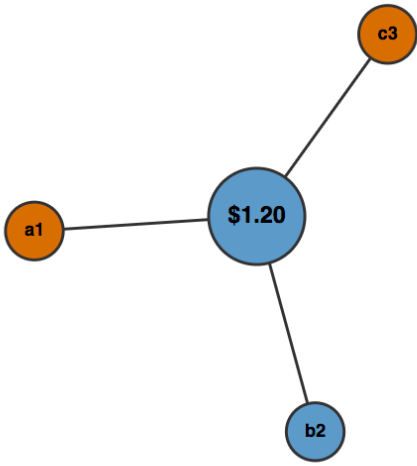
**Tutorial 5/14**

For example, if you **choose 'A'**, you keep your money for yourself.

**Next**

## Data S1. Instruction and tutorials, Related to Figure 1. (cont.)

The game will start in: 05:58



**Tutorial 6/14**

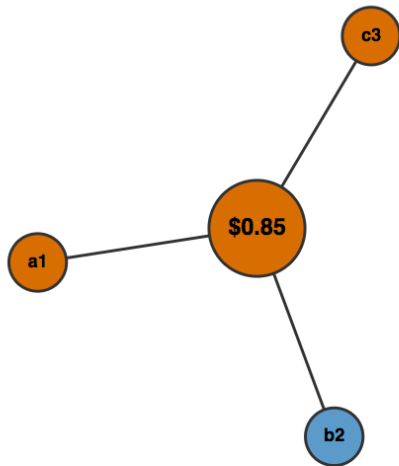
After you make your choice, you will be informed of your partners' choices by their node color.

In this example, 1 partner **chose 'A'** and kept money for themselves. On the other hand, 2 partners **chose 'B'** and paid \$0.05 each to contribute a total of \$0.20 to you.

Since you **chose 'A'**, your partners did not get any earnings from you. You earned money from your partners without losing any of yours.

**Next**

The game will start in: 05:47



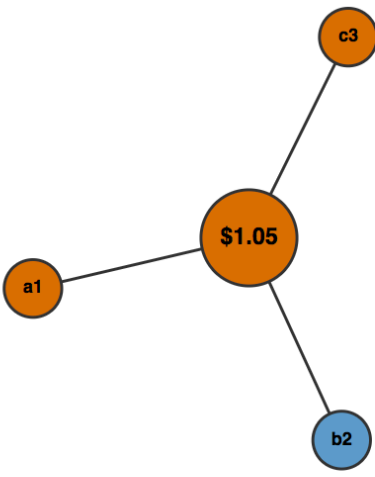
**Tutorial 7/14**

if you **choose 'B'** with three game partners, you pay \$0.05 to contribute \$0.10 to each of them. You lose \$0.15 in this example.

**Next**

Data S1. Instruction and tutorials, Related to Figure 1. (cont.)

The game will start in: 05:38



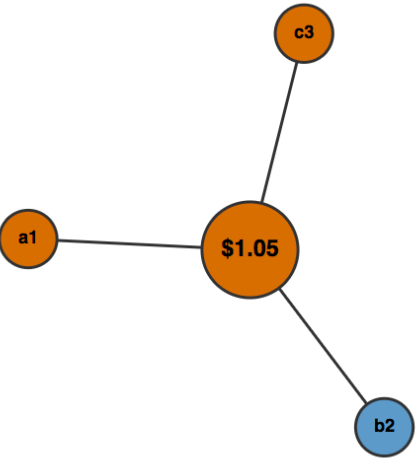
**Tutorial 8/14**

Since two of your partners also **chose 'B'**, you earned \$0.20 from them in total.

After you learn your partners' choices, you will move on to Step 2.

**Next**

The game will start in: 05:30



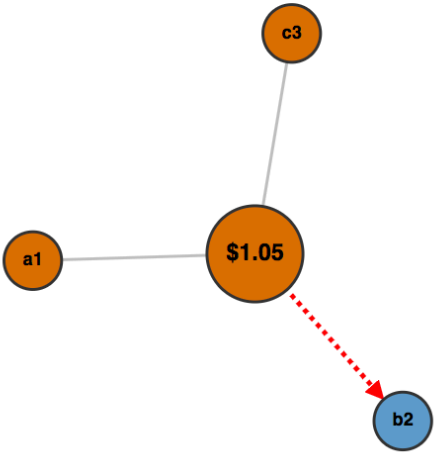
**Tutorial 9/14**

In Step 2, you may choose to make or break connections with other players. To help you make an informed decision, we will show you the player's last choice: 'A' (**keep money for self**) or 'B' (**give money to others**).

**Next**

## Data S1. Instruction and tutorials, Related to Figure 1. (cont.)

The game will start in: 05:22

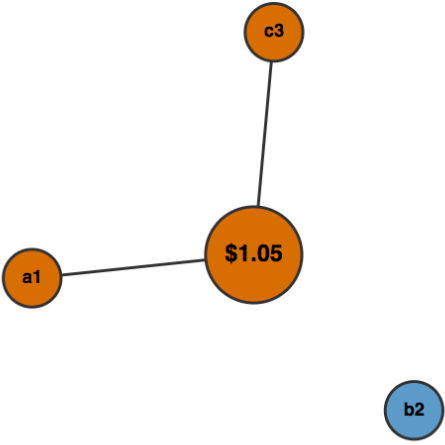


Tutorial 10/14

You may be asked if you want to **cut the connection** with your current partner.

**Next**

The game will start in: 05:14



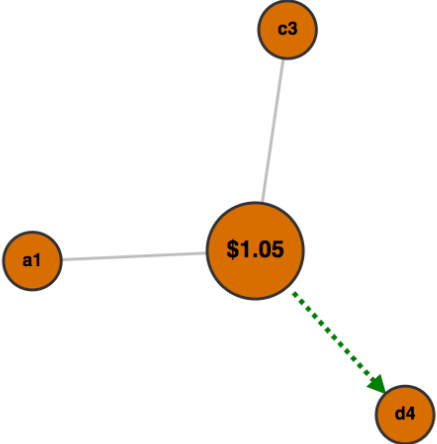
Tutorial 11/14

When you **cut the connection**, you won't play with the player in future rounds.

**Next**

## Data S1. Instruction and tutorials, Related to Figure 1. (cont.)

The game will start in: 05:05

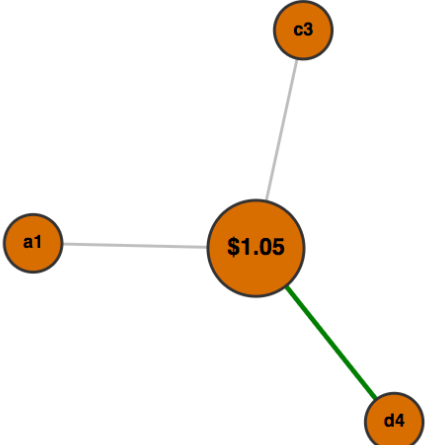


Tutorial 12/14

You may be also asked if you would like to **make a connection** with a new player.

Next

The game will start in: 04:57



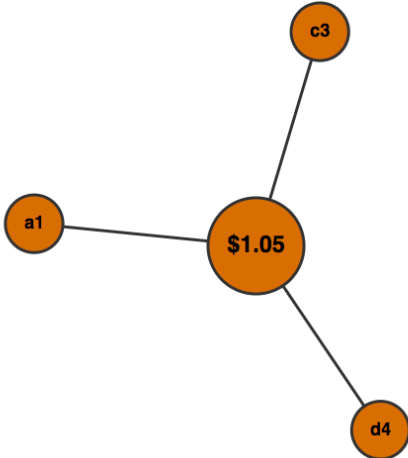
Tutorial 13/14

When you **make a connection**, you will play with the new partner in future rounds.

Next

## Data S1. Instruction and tutorials, Related to Figure 1. (cont.)

The game will start in: 04:48



**Tutorial 14/14**


In the next round, you will choose whether to give money to your current partners with knowledge of their last choice. You will repeat the sequence for several rounds.

**Note that you will be removed from the game if other players are waiting on you to make a decision for longer than 1 minute.**

When you complete the game, you will be paid your final earnings plus \$1.00 as completion bonus.

**Next**

The game will start in: 04:39



**You have completed the tutorial.**

Now you are ready to play a practice game.

You will play 2 practice rounds. Your partners are all programmed "bots" in the practice rounds.

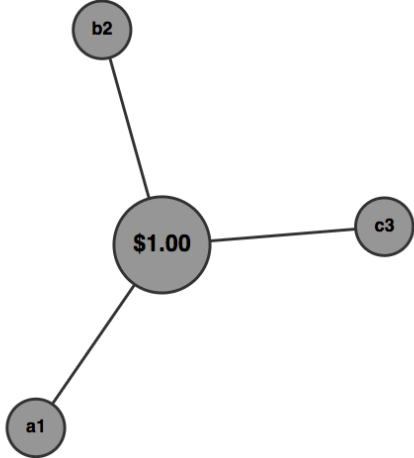
**The results of this practice game will not change your bonus.**

Click 'Start Practice' to begin.

**Start Practice**

**Data S1. Instruction and tutorials, Related to Figure 1. (cont.)**

The game will start in: 04:29



**Step 1 (practice game)**

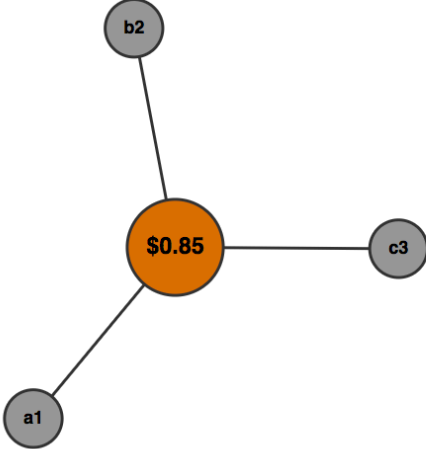
*The result of these rounds will not affect your bonus.*

**Your earnings: \$1.00**

- If you **choose 'A'**, you keep your money for yourself.
- If you **choose 'B'**, you give \$0.05 to each partner; each of your partners earns \$0.10.

**A(-\$0.00)**   **B(-\$0.15)**

The game will start in: 04:19



**Step 1 (practice game)**

*The result of these rounds will not affect your bonus.*

**Your earnings: \$0.85**

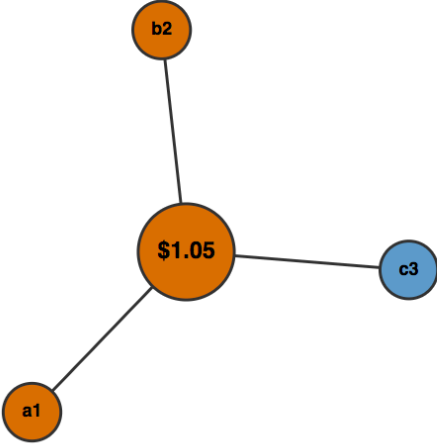
**You chose 'B'**. You paid \$0.15 total to contribute \$0.10 to each partner.

**Next**



**Data S1. Instruction and tutorials, Related to Figure 1. (cont.)**

The game will start in: 04:10



A central orange circle labeled "\$1.05" is connected by solid lines to three other nodes: "a1" (orange), "b2" (orange), and "c3" (blue).

**Step 1 (practice game)**

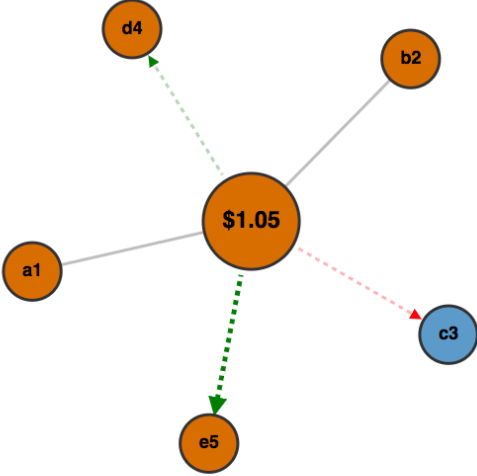
*The result of these rounds will not affect your bonus.*

**Your earnings: \$1.05**

2 of your partners paid \$0.05 each; you earned \$0.20 from them.

**Next**

The game will start in: 04:02



A central orange circle labeled "\$1.05" is connected to five nodes: "a1" (orange), "b2" (orange), "c3" (blue), "d4" (orange), and "e5" (orange). Solid lines connect the central node to "a1", "b2", and "c3". Dotted green lines with arrowheads at the end connect the central node to "d4" and "e5". A dotted red line with an arrowhead at the end connects the central node to "c3".

**Step 2 (practice game)**

*The result of these rounds will not affect your bonus.*

**Your earnings: \$1.05**

You are not currently connected to this player; you can choose to **make a connection**.

**e5** This player **chose 'B'** (give money to others) in the last step.

Do you want to **make a connection** with this player?

**Make** **Do not make**

Data S1. Instruction and tutorials, Related to Figure 1. (cont.)

The game will start in: 03:53

**Step 2 (practice game)**

The result of these rounds will not affect your bonus.

**Your earnings: \$1.05**

You are not currently connected to this player; you can choose to **make a connection**.

**d4** This player chose 'B' (give money to others) in the last step.

Do you want to **make a connection** with this player?

**Make** **Do not make**

The game will start in: 03:45

**Step 2 (practice game)**

The result of these rounds will not affect your bonus.

**Your earnings: \$1.05**

You are currently connected to this player; you can choose to **cut the connection**.

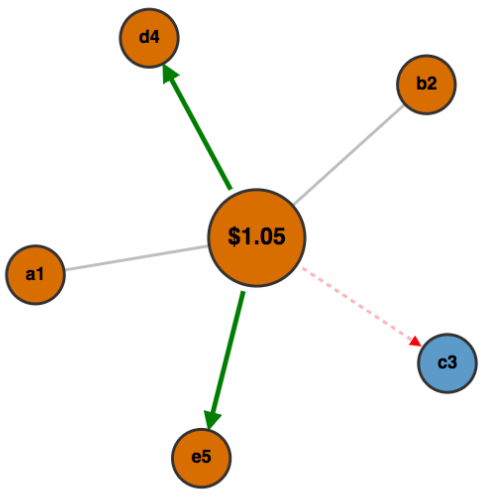
**c3** This player chose 'A' (keep money for self) in the last step.

Do you want to **cut the connection** with this player?

**Cut** **Do not cut**

## Data S1. Instruction and tutorials, Related to Figure 1. (cont.)

The game will start in: 03:27



**Step 2 (practice game)**

*The result of these rounds will not affect your bonus.*

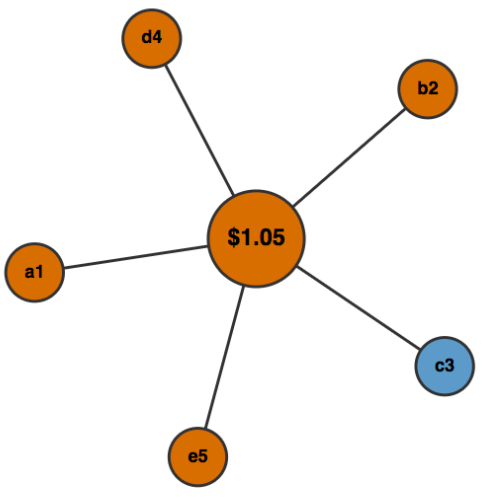
**Your earnings: \$1.05**

This round:

- you made 2 connection(s) with player(s)

**Next**

The game will start in: 03:19



**Step 1 (practice game)**

*The result of these rounds will not affect your bonus.*

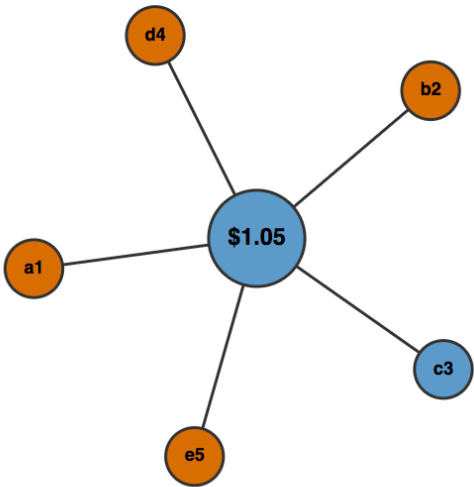
**Your earnings: \$1.05**

- If you **choose 'A'**, you keep your money for yourself.
- If you **choose 'B'**, you give \$0.05 to each partner; each of your partners earns \$0.10.

**A(-\$0.00)**   **B (-\$0.25)**

**Data S1. Instruction and tutorials, Related to Figure 1. (cont.)**

The game will start in: 03:10



The diagram shows a central blue circle labeled "\$1.05" connected to five other circles: a1 (orange), d4 (orange), b2 (orange), c3 (blue), and e5 (orange).

**Step 1 (practice game)**

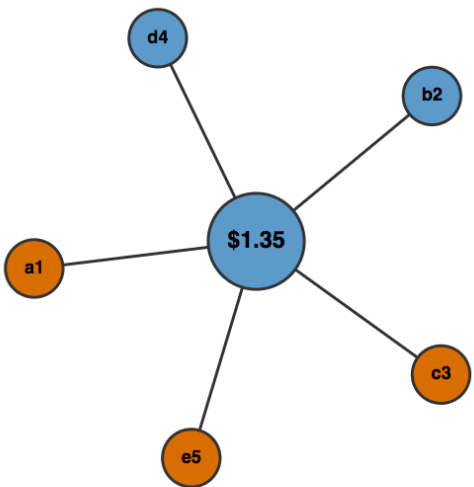
*The result of these rounds will not affect your bonus.*

**Your earnings: \$1.05**

**You chose 'A'.** You paid \$0.00 total to contribute \$0.00 to each partner.

**Next**

The game will start in: 03:02



The diagram shows a central blue circle labeled "\$1.35" connected to five other circles: a1 (orange), d4 (blue), b2 (blue), c3 (orange), and e5 (orange).

**Step 1 (practice game)**

*The result of these rounds will not affect your bonus.*

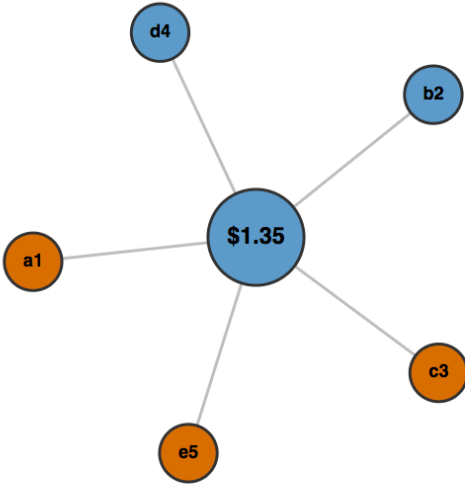
**Your earnings: \$1.35**

3 of your partners paid \$0.05 each; you earned \$0.30 from them.

**Next**

## Data S1. Instruction and tutorials, Related to Figure 1. (cont.)

The game will start in: 02:55



**Step 2 (practice game)**

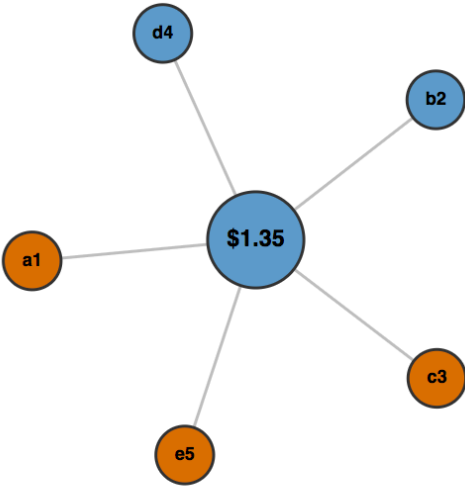
*The result of these rounds will not affect your bonus.*

**Your earnings: \$1.35**

You have no options to make or cut a connection in this round.

**Next**

The game will start in: 02:46



**Step 2 (practice game)**

*The result of these rounds will not affect your bonus.*

**Your earnings: \$1.35**


This round:

- There were no changes to your connections this round.

**Next**

## Data S1. Instruction and tutorials, Related to Figure 1. (cont.)

The game will start in: 02:36




You finished the practice game.

Now that you have completed the practice, please answer the comprehension questions. For each question, you can only choose one answer.

**If you answer all three questions correctly, you will be able to join the game and earn a bonus.**

**Next**

The game will start in: 02:27



Test 1/3

Please choose the best answer.

**Q1. When will the game end?**

A1. After several rounds, but you will not be informed of the number of rounds.

A2. After you play one round.

A3. When your score is less than 0.

**A1**   **A2**   **A3**

## Data S1. Instruction and tutorials, Related to Figure 1. (cont.)

The game will start in: 02:16


Test 2/3

Please choose the best answer.

**Q2. How will a connection be cut?**

A1. Either you or the other player agree to cut.  
A2. You cannot cut a connection.  
A3. Both you and the other player agree to cut.

A1 A2 A3



The game will start in: 02:07

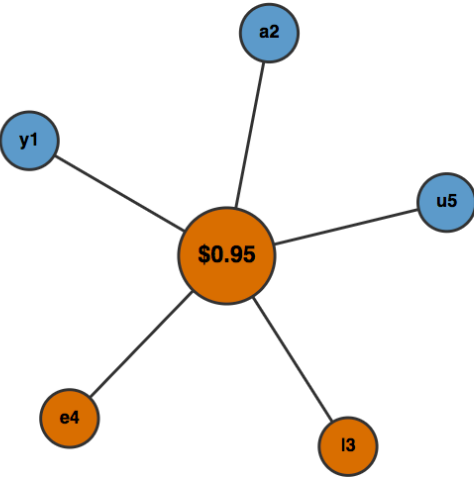
Test 3/3

Please choose the best answer.

**Q3. Which sentence properly explains the situation to the left?**  
(blue (A): keep money for self; orange (B): pay \$0.05 to give \$0.10 to each partner)

A1. You paid \$0.25 and received \$0.20 in total in this round.  
A2. The blue-colored players earned nothing in this round.  
A3. There are only five players in the entire network.

A1 A2 A3



**Data S1. Instruction and tutorials, Related to Figure 1. (cont.)**


The game will start in: 01:57

You are now ready to join the game.

Please wait for the other players to complete the tutorial.

**When the timer at the top elapses, a 'Ready' button will appear. Please click the Ready button to begin.**

**If you don't see a 'Ready' button after the timer elapses, please refresh your browser.**




You are now ready to join the game.

Please wait for the other players to complete the tutorial.

**When the timer at the top elapses, a 'Ready' button will appear. Please click the Ready button to begin.**

**If you don't see a 'Ready' button after the timer elapses, please refresh your browser.**

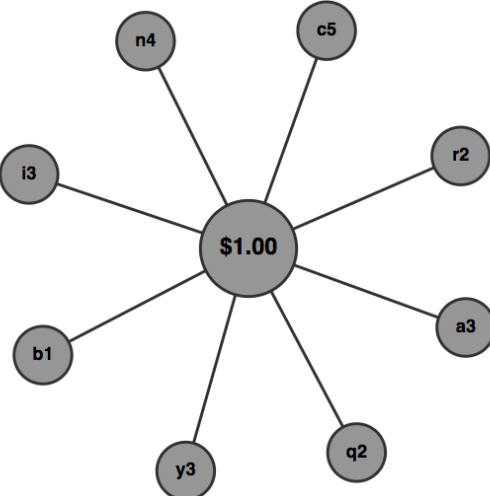
**Ready**





### Data S1. Instruction and tutorials, Related to Figure 1. (cont.)

Sample screenshots of the real games (main experiments; bot-invisible condition):

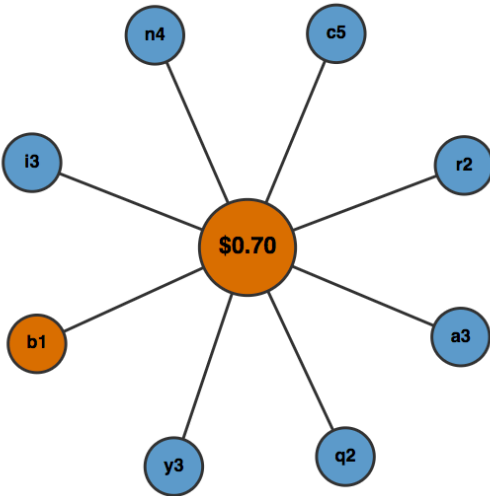


**Step 1**

**Your earnings: \$1.00**

- If you **choose 'A'**, you keep your money for yourself.
- If you **choose 'B'**, you give \$0.05 to each partner; each of your partners earns \$0.10.

**A (-\$0.00)**   **B (-\$0.40)**



**Step 1**

**Your earnings: \$0.70**

1 of your partners paid \$0.05 each; you earned \$0.10 from them.

**Next**

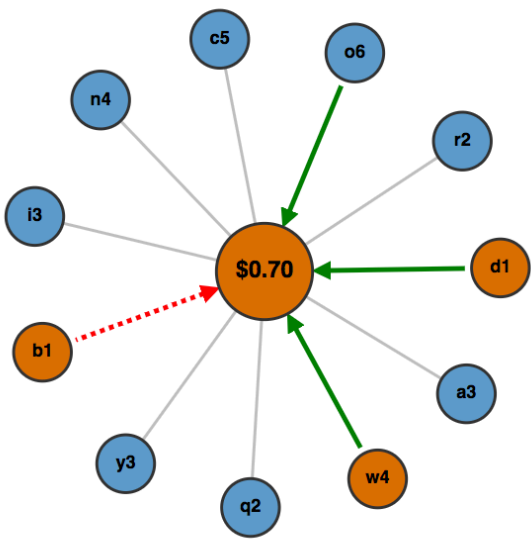
Data S1. Instruction and tutorials, Related to Figure 1. (cont.)

### Step 2

**Your earnings: \$0.70**  
This round:

- 3 player(s) made their connection(s) with you
- 1 player(s) broke their connection(s) with you

**Next**



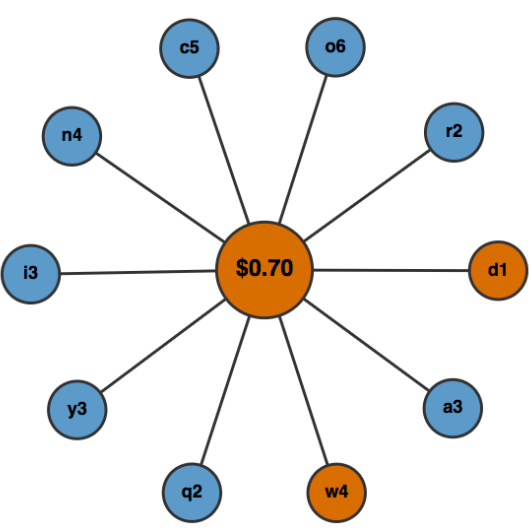
The diagram shows a central orange node labeled '\$0.70' connected to ten peripheral nodes: n4, c5, o6, r2, d1, a3, w4, q2, y3, and i3. Green arrows point from o6, d1, and w4 towards the central node. A red dashed arrow points from b1 towards the central node. The nodes n4, c5, o6, r2, a3, y3, q2, and i3 are blue, while d1, w4, and b1 are orange.

### Step 1

**Your earnings: \$0.70**

- If you **choose 'A'**, you keep your money for yourself.
- If you **choose 'B'**, you give \$0.05 to each partner; each of your partners earns \$0.10.

**A(-\$0.00)**   **B (-\$0.50)**



The diagram shows a central orange node labeled '\$0.70' connected to ten peripheral nodes: n4, c5, o6, r2, d1, a3, w4, q2, y3, and i3. All nodes are connected to the central node by simple lines. The nodes n4, c5, o6, r2, a3, y3, q2, and i3 are blue, while d1, w4, and b1 are orange.

Data S1. Instruction and tutorials, Related to Figure 1. (cont.)

A central blue node labeled '\$1.30' is connected to eight peripheral nodes: c5, o6, r2, a3, q2, i3, n4, and e8. Node e8 is orange and has a green dashed arrow pointing towards the central node.

### Step 2

**Your earnings: \$1.30**  
You are not currently connected to this player; you can choose to **make a connection**.

This player chose 'B'(give money to others) in the last step  
in response to 1 of their 6 neighbors choosing 'B'.  
Do you want to **make a connection** with this player?

A central orange node labeled '\$2.55' is connected to eight peripheral nodes: e8, o6, r2, a3, q2, y3, i3, and n4. Node r2 is blue and has a red dashed arrow pointing towards the central node. Node y3 is orange and has a green dashed arrow pointing towards the central node.

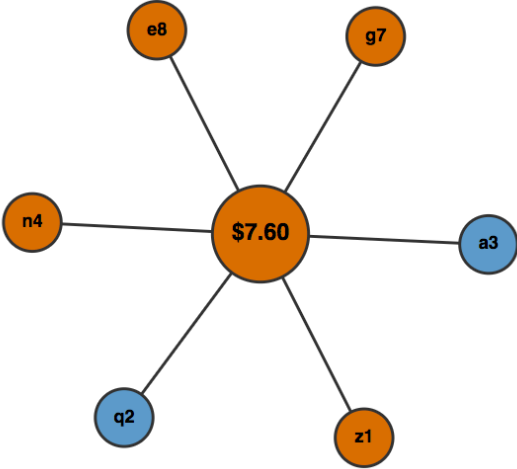
### Step 2

**Your earnings: \$2.55**  
You are currently connected to this player; you can choose to **cut the connection**.

This player chose 'A'(keep money for self) in the last step  
in response to 5 of their 6 neighbors choosing 'B'.  
Do you want to **cut the connection** with this player?

**Data S1. Instruction and tutorials, Related to Figure 1. (cont.)**

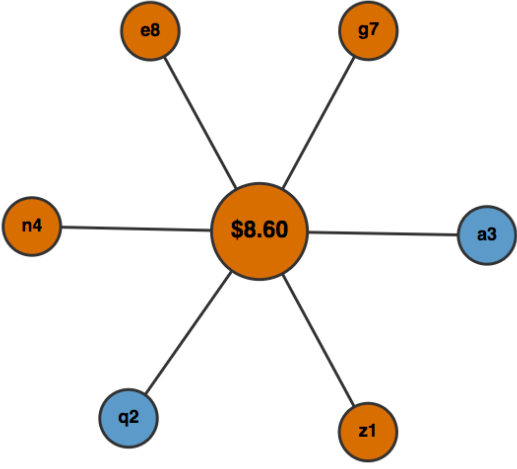
A session finished after players repeated the cooperation and rewiring steps 30 times.



You completed the game.

You will get \$7.60 from your game score, in addition to the completion bonus \$1.00. **Your total bonus is \$8.60.**

Next



Thank you for playing!

Please click the 'Submit HIT' button to submit your HIT after you answer the following questions.

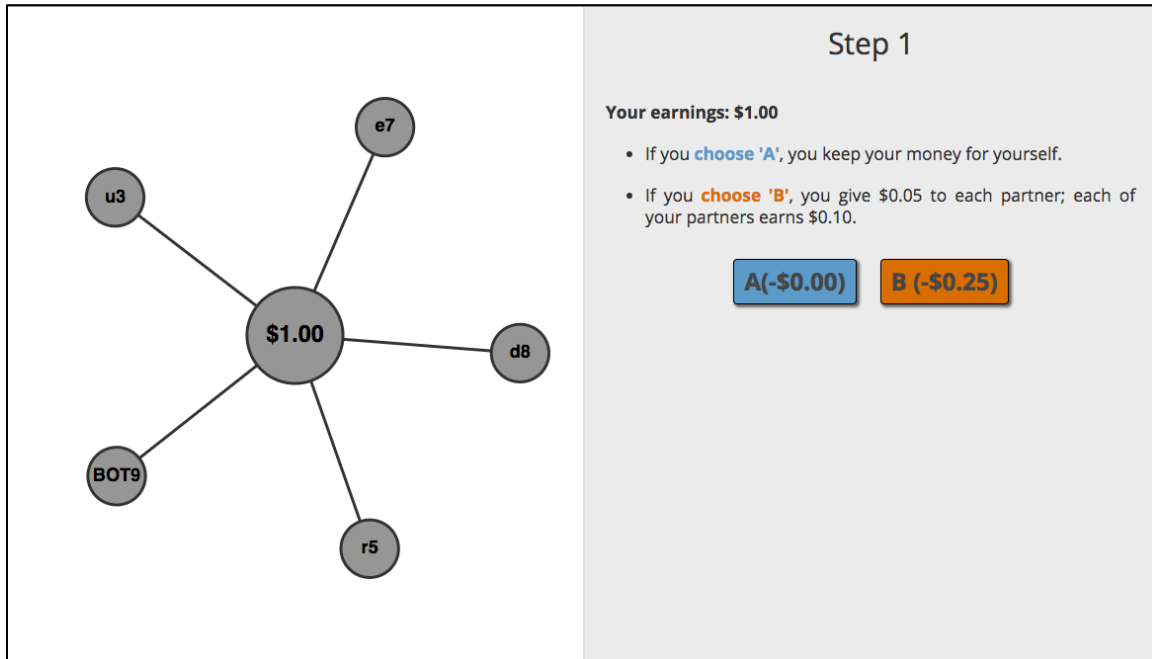
1. What strategy did you employ in the game?

2. How did you feel about your partners?

Submit HIT

**Data S1. Instruction and tutorials, Related to Figure 1. (cont.)**

Sample screenshots of the real games (supplementary experiments; bot-visible condition):



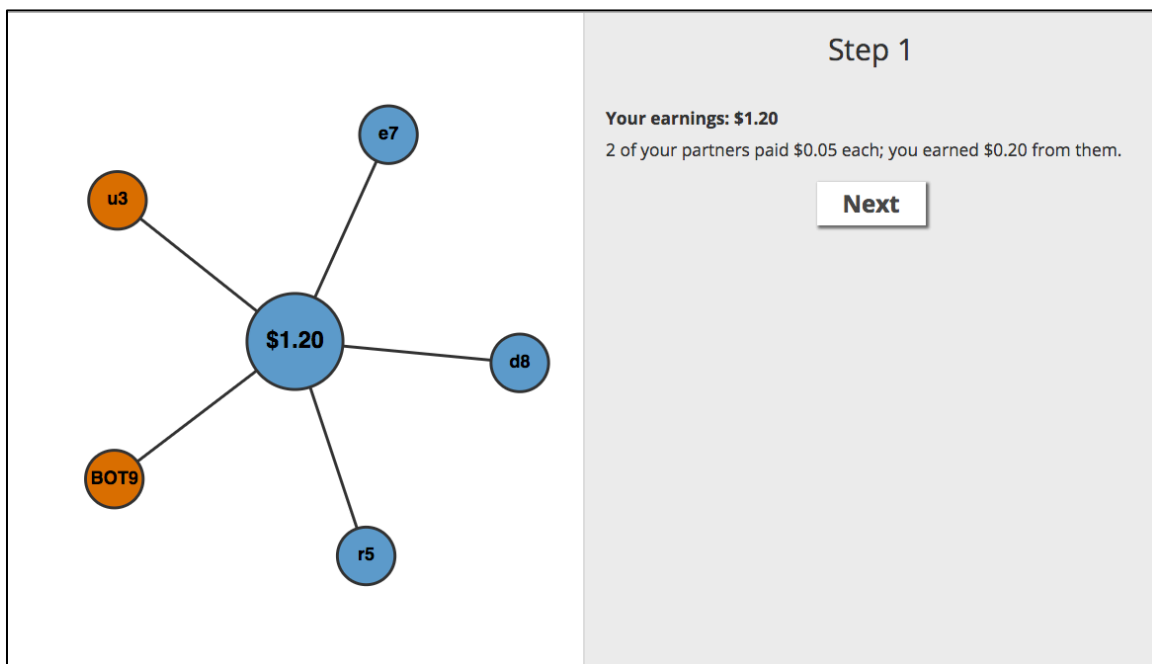
**Step 1**

**Your earnings: \$1.00**

- If you **choose 'A'**, you keep your money for yourself.
- If you **choose 'B'**, you give \$0.05 to each partner; each of your partners earns \$0.10.

**A(-\$0.00)** **B(-\$0.25)**

The diagram shows a central grey node labeled "\$1.00" connected to five other grey nodes: "u3", "e7", "d8", "r5", and "BOT9".



**Step 1**

**Your earnings: \$1.20**

2 of your partners paid \$0.05 each; you earned \$0.20 from them.

**Next**

The diagram shows a central blue node labeled "\$1.20" connected to five other nodes: "u3", "e7", "d8", "r5", and "BOT9". The nodes "u3", "e7", "d8", and "r5" are blue, while "BOT9" is orange.

Data S1. Instruction and tutorials, Related to Figure 1. (cont.)

```
graph TD; C((\$1.20)) --- U3((u3)); C --- BOT9((BOT9)); C --- E7((e7)); C --- R5((r5)); C -.-> D8((d8));
```

### Step 2

**Your earnings: \$1.20**

You are currently connected to this player; you can choose to **cut the connection**.

**d8** This player chose 'A'(keep money for self) in the last step in response to 3 of their 7 neighbors choosing 'B'.

Do you want to **cut the connection** with this player?

**BOT9 recommends that you cut the connection.**

```
graph TD; C((\$1.20)) --- U3((u3)); C --- BOT9((BOT9)); C --- E7((e7)); C --- R5((r5)); E7 -.-> C; R5 -.-> C; C -.-> D8((d8));
```

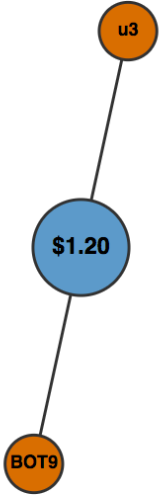
### Step 2

**Your earnings: \$1.20**

This round:

- you broke 1 connection(s) with player(s)
- 2 player(s) broke their connection(s) with you

Data S1. Instruction and tutorials, Related to Figure 1. (cont.)



Step 1

Your earnings: \$1.20

- If you **choose 'A'**, you keep your money for yourself.
- If you **choose 'B'**, you give \$0.05 to each partner; each of your partners earns \$0.10.

**A(-\$0.00)**   **B (-\$0.10)**

## 2. Transparent methods

### *Recruitment procedure*

A total of 1,024 subjects ( $N_{subject}=896$  for Experiment 1;  $N_{subject}=128$  for Experiment 2) participated in our incentivized decision-making game experiments, always in groups of 16 people ( $N_{session}=64$  sessions). Subjects were recruited using Amazon Mechanical Turk (MTurk) via our breadboard software platform (which we have made available at [breadboard.yale.edu](http://breadboard.yale.edu)). MTurk is an online labor market in which employers contract with workers to complete short tasks for relatively small amounts of money. Many studies have demonstrated the validity of behavioral experiment data gathered using MTurk (Rand, 2012; Thomas and Clifford, 2017). Behaviors of MTurk subjects in stylized economic games are correlated with their actual behaviors in a real-world situation (Peysakhovich et al., 2014). MTurk subjects often receive a baseline payment, plus an additional bonus depending on their performance. For this study, subjects received \$2.00 when they completed the tutorial to join a game and \$1.00 when they completed the entire game. In addition, subjects also received a bonus payment based on their own earnings during the networked cooperation game, which averaged \$9.50.

### *Experimental setup*

Our participants interacted anonymously using *breadboard* and playing in a browser window. We prohibited subjects from participating in more than one session of the experiment by using the unique identifications for each subject on MTurk. All the subjects were informed about the use of their behavioral data for research purpose upon



enrollment in the experiment. The experiments were conducted from August to November 2018.

Subjects were placed in a group of humans with a size of 16 and arranged in a social network with an Erdős-Rényi random graph configuration in which 30% of ties were present, on average. In a bot-integrated session of Experiment 1, each subject additionally received one connection with a bot. Subjects were therefore initially connected to an average 4.41 (s.d. = 1.76) human neighbors and 1 bot neighbor (i.e., average 5.41 neighbors in total). In Experiment 2, we created a random network with 16 human subjects, as above, and then added 1 bot having 5 ties to the network, which resulted in an average of 5.13 neighbors in the whole group at the onset (sometimes including the bot). Subjects could identify each neighbor by a name label that was randomly generated with a number and a letter (such as “j2” and “y9”; see Data S1).

The subjects played a Public Goods game lasting 30 rounds with their network neighbors (which consisted of other human subjects and possibly bots). At the beginning of the game, human subjects received \$1.00 as their initial endowment. In each round, all the subjects chose whether to cooperate, by reducing their own endowment \$0.05 per neighbor in order to increase the endowment of all neighbors by \$0.10 each, or to defect, by paying no cost and providing no benefits. Subjects made the same choice with all their neighbors. When a subject did not have any neighbors in this step, the subject was not given the cooperation choice at the round.

After making their cooperation choice, subjects were informed of the choices made by their neighbors. Then, subjects sometimes had the opportunity to change their neighbors by making or breaking ties (“tie-rewiring” options). Specifically, 5% of all pairs of human subjects were chosen at random in each round and given the opportunity to rewire their networks. If a tie already existed between the two subjects, then one of the two was picked at random to be allowed to choose whether to voluntarily break the tie with the other; if a tie did not already exist between the two, a randomly selected subject from the pair was given the option to form a tie. When making this decision, subjects were aware of whether the person to whom they might disconnect or connect had cooperated or defected in the past round. In addition, they were also informed how many total neighbors and cooperative ones the focal person had in his or her immediate environment.

At any point during the game, if a subject was inactive for 15 seconds, the subject was warned about being dropped. If they still remained inactive after 15 seconds, they were dropped. Since dropping subjects changed the network structure, we calculated all the network metrics in each session of 30 rounds and used the average for each subject. The dropped subjects were prohibited from joining another session of this experiment.

Within this basic setup, we introduced 16 bots into the network of 16 human subjects (except for the control sessions without bots) in Experiment 1. Each bot had only one tie and connected with a different subject (that is, each subject had a bot among their neighbors) at the beginning of a game. We used the artifice of single-tie bots so as to fix

the amount of intervention across sessions and treatments to 16 total ties with bots at all times and in all treatments. These single-link bots keep the same amount of intervention (in terms of ties and money) across sessions and treatments. Moreover, this set-up made all the subjects interact with one cooperative (or Tit-for-Tat) bot over rounds. Thus, we can be sure that whether cooperation collapses is unrelated to any heterogeneity of bot influence across groups. If only some of the subjects had bots at the outset, the ineffectiveness (or effectiveness) could come from the possibly biased characteristics of subjects that bots attached to by chance. But here, since every subject equally has one connection with a bot, this is not an issue.

Bots always chose cooperation in the game (except for the sessions of Tit-for-Tat bots); that is, they gave the same amount of cooperation benefit into a social system ( $\$1.60 = 16 \times \$0.10$  per round) to the subjects who connected with them. Bots never connected with each other.

Subjects were not informed that there were bots in the game (except in the extra condition of bot visibility). In their local view, subjects could identify every bot as a neighbor in the same manner as other subjects using the name labeling system.

In Experiment 2, we explored the possibility of a minimal intervention based on the results of Experiment 1. In this experiment, we added 1 bot having 5 connections to a network of 16 human subjects. In contrast of Experiment 1, some subjects had a connection with the bot and the others did not. Like Experiment 1, bots always chose

cooperation in the game so that they gave a total of \$0.50 to the network of humans at each round ( $\$0.50 = 5 \times \$0.10$ ).

### *Statistical analysis*

Analyzing the data from our experiment requires more than an analysis of the average final values across treatment groups (which leverages the randomization of the experiment). For instance, the average values by round do not represent directly the slopes of the change over a session. Moreover, ratio data represented as the cooperation rate, which is limited to be between 0 and 1, does not come from a population that is normally distributed. In addition, multiple observations from the same subject and observations from multiple subjects within the same session are not independent. Thus, we need to deal with the nested structure of errors in our statistical analysis.

Hence, we used a statistical analysis based on a GLMM involving logistic regression with nested random effects. GLMM estimates coefficients in the linear predictor and random effects, which comes from individual differences, at the same time, using maximum-likelihood methods. All analysis was performed using R version 3.6.0.

To be concrete, here is how the model was implemented to get the statistical results on strategy selection in Fig. 2: let  $p_{t,i,k}$  denote the probability of player  $i$  selecting cooperation at round  $t$  in a session  $k$ ; let  $I_{k \in A}$  be the vector of dummy variables that indicates whether the session  $k$  belongs to an experimental treatment  $A$ ; let  $\varepsilon_k$  be the

random effects of the session  $k$ ; let  $\varepsilon_{i|k}$  be the random effects of player  $i$  nested within the session  $k$ ; and let  $\varepsilon_{t,i,k}$  be the error. Thus, we have

$$p_{t,i,k} = 1/1\{1 + \exp(-z_{t,i,k})\} \quad (1)$$

$$z_{t,i,k} = \beta_0 + \beta_1 I_{k \in A} t + \varepsilon_k + \varepsilon_{i|k} + \varepsilon_{t,i,k} \quad (2)$$

We used a logistic function as the link function for the statistical modeling. The random effects  $\varepsilon_k$  and  $\varepsilon_{i|k}$  are approximated by the normal distribution with mean value zero. Fig. 2 shows the estimated  $\beta_1$  for each treatment.

To compare the slope of disengaged intervention with that of each other treatment, we modified the equation (2), so that

$$z_{t,i,k} = \beta_0 + (\beta_1 + \beta_2 I_{k \in A}^*) t + \varepsilon_k + \varepsilon_{i|k} + \varepsilon_{t,i,k} \quad (3)$$

where  $I_{k \in A}^*$  is the vector of dummy variables that indicates whether the session  $k$  belongs to an experimental treatment  $A$  instead of the disengaged tie-management. In other words, the statistical model (3) uses the disengaged tie-management as the reference category.

Table S1 shows the estimated  $\beta_0$ ,  $\beta_1$ , and  $\beta_2$  and their standard errors.

We may assume that each subject makes their cooperation decision based on their neighborhood environment. We modeled the local influence on cooperation decision-making based on GLMM, so that

$$p_{t,i,k} = 1/1\{1 + \exp(-z_{t,i,k})\} \quad (4)$$

$$z_{t,i,k} = \beta_0 + \beta_X X_{t,i,k} + \beta_t t + \varepsilon_k + \varepsilon_{i|k} + \varepsilon_{t,i,k} \quad (5)$$

In model (5), the covariate  $X_{t,i,k}$  is the vector of the number of the neighbors of subject  $i$  at round  $t$  in the session  $k$   $x_{t,i,k}$ , the rate of cooperators in the neighbors of subject  $i$  at

round  $t$  in the session  $k$   $r_{t,i,k}^C$ , and the number of cooperators in the neighbors of subject  $i$  at round  $t$  in the session  $k$   $x_{t,i,k}^C$ .

In the estimation, we found that the random effect for sessions,  $\varepsilon_k$ , was nearly zero in our experiment data. Considering the issue of calculation convergence, we removed the session-level random effect from model (5):

$$z_{t,i,k} = \beta_0 + \beta_X X_{t,i,k} + \beta_t t + \varepsilon_i + \varepsilon_{t,i,k} \quad (6)$$

The model still has the random effects for individuals,  $\varepsilon_i$ . We also confirmed the robustness of the environment effects with the model controlling the status-quo bias (i.e., lagged cooperation), so that

$$z_{t,i,k} = \beta_0 + \beta_X X_{t,i,k} + \beta_t t + \beta_a a_{t-1,i,k} + \varepsilon_i + \varepsilon_{t,i,k} \quad (7)$$

where the covariate  $a_{t-1,i,k}$  is a binary variable of whether the subject  $i$  chose cooperation at round  $t-1$ . Table S2 shows all the estimated coefficients in models (6) and (7). We calculated the cooperation probabilities in Figure 3A by placing the estimated coefficients in model (6) and then model (4).

### **3. Supplementary references**

Peysakhovich, A., Nowak, M.A., and Rand, D.G. (2014). Humans display a “cooperative phenotype” that is domain general and temporally stable. *Nature Communications* 5. doi: 10.1038/ncomms5939.

Rand, D.G. (2012). The promise of Mechanical Turk How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology* 299, 172–179.

Thomas, K.A., and Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior* 77, 184–197.