A stranger in a strange land: Promises and identity*

Gary Charness, Giovanni Di Bartolomeo, and Stefano Papa

April 23, 2022

Abstract. Social identity and communication are topics of increasing interest in management science. One's social identity tends to lead one to favor those belonging to one's group; this in-group bias may lead to problematic relationships. At the same time, communication has been found to have beneficial social consequences in controlled laboratory experiments. An important question is whether communication, by signaling a meeting of the minds, can improve trust and therefore outcomes between out-group members. We construct a simple weak mechanism of group favoritism that does in fact show in-group favoritism. When both paired individuals, one of whom will become the dictator, promise to make the pro-social dictator choice if they become dictator, favorable behavior is much more likely in all cases. But there is an intriguing pattern across group membership concerning the degree of improvement: Without mutual promises, people make more favorable choices for in-group members. Interestingly, this gap is eliminated by such promises. In this sense, strangers become partners.

***JEL* codes:** A13, C91, D03, D64, D90.

Keywords: Social identity, in-group bias, communication, exogenous variation.

* We thank Giuseppe Attanasi, Martin Dufwenberg, and Kiryl Khalmetski for valuable comments.
Charness, University of California, Santa Barbara, charness@econ.ucsb.edu; Di Bartolomeo, Sapienza University of Rome and University of Antwerp, giovanni.dibartolomeo@uniroma1.it; Papa, University of Rome Tor Vergata, stefano.papa@uniroma2.it

1. Introduction

Social identity and communication are topics of increasing interest in the management literature, involving perspectives related to strategy, entrepreneurship, innovation, technology, and organizations.[1] Experimental literature has shown that individuals tend to prefer those who belong to their own group (in-group bias). Often even artificial constructions of affiliation to a group are enough to produce in-group favoritism effects.[2] At the same time, a series of laboratory experiments have shown that costless and non-binding communication (*cheap talk*) can be effective in removing barriers to social efficiency. Might it also improve outcomes when people in relative out-groups are interacting, thereby overcoming a reluctance to trust strangers? We design and conduct a laboratory experiment to shed light on this question.

The issue of one's identity has become a prominent feature of contemporary society, particularly in the political arena. In recent years, identity has been used more and more as a wedge to separate subgroups. It is important to understand the ramifications of identity, both to limit the negative consequences and to be able to use one's sense of identity as a positive force in our world. One insight from social-identity theory (Tajfel and Turner, 1979) is that people derive self-esteem from group membership and adopt behaviors that are consistent with the norms and stereotypes associated with that group identity. Akerlof and Kranton (2000), Charness, Rigotti, and Rustichini (2007), Chen and Li (2009), and Chen and Chen (2011) introduce identity issues into the experimental literature, and Ockenfels and Werner (2014) and Ciccarone, Di Bartolomeo, and Papa (2020) extend the issue to explore the relationship between in-group favoritism and beliefs.[3]

According to the Tajfel and Turner (1979) theory, social identity has three components: categorization, identification, and comparison. Categorization involves labeling people as, for

[1] Some (non-exhaustive) examples are related to male stereotypes in talent allocation (Coffman, 2014; Niessen-Ruenzi and Ruenzi, 2019; Del Carpio and Guadalupe, 2021), religious identity and mutual fund risk-taking behaviors (Shu, Sulaeman, and Yeung, 2012), product user identity and consumption choices (Bagozzi and Dholakia 2006; Dahl, Fuchs, and Schreierm, 2015), human identity vs. machine in cheating (Cohn, Gesche, and Maréchal, 2021), and identity based on common interests vs. competition in a small R&D firm (Reagans, 2005).

[2] Even minimal group assignments – or less, see Charness and Holder, 2018 – can affect behavior (see also Chen and Li, 2009).

[3] Ockenfels and Werner (2014) showed that in-group favoritism may be belief dependent when dictators can strategically manipulate recipients' beliefs. Refining Ockenfels and Werner (2014), Ciccarone, Di Bartolomeo and Papa (2020) provide support to the idea that people have a general intrinsic preference for group members.

example, Catholic, female, or Black, thereby implicitly defining them. Similarly, our self-image is associated with the categories to which we belong. Identification reflects how we associate ourselves with certain groups, where in-groups are groups with which we identify, and outgroups are ones with which we don't identify. The third component, comparison, is the process by which we compare our groups with other groups. There is little research on which facet of one's identity comes to the fore (although see).[4] Nevertheless, there is evidence (e.g., Shih et al., 1999) that this is sensitive to the environment and to cues. To the extent that this is true, there is scope for encouraging a more favorable identity to emerge.

Gender identity is particularly important in the management realm (e.g., Coffman, 2014; Niessen-Ruenzi and Ruenzi, 2019; Del Carpio and Guadalupe, 2021). For example, the strong male stereotype associated with some careers inhibit talented females from applying for such positions; this leads to production inefficiencies and mismatching for organizations. Del Carpio and Guadalupe (2021) study whether the effects of different recruiting strategies for tech positions might counterbalance strong male stereotypes. By using communication, in the form of informational messages, application rates are substantially increased (including those of candidates at the top of the cognitive skill distribution); however, communication also introduces negative selection on cognitive skills, implying a higher cost of screening.

The decision environment and the available communication technology are instrumental in determining the effect of this communication on choices and behavior. In some cases (e.g., Cooper, DeJong, Forsythe, and Ross, 1989; Charness, 2000a), sending a simple message stating "I intend to play [A or B]" leads to a great increase in social efficiency in coordination games (multiple equilibria). However, this message was completely ineffective in a prisoner's dilemma game, where there is only one socially-inefficient equilibrium. Brandts, Charness, and Ellman (2016) and Charness, Feri, Melendez, and Sutter (2021) find very large beneficial effects from chat in games, but no effect in these same games from simple, check-a-box messages.[5]

Closer to our setting, Charness and Dufwenberg (2006) find that self-generated, free-form written promises were especially effective in generating social efficiency: The outcomes

[4] Exceptions include Chen, Li, Liu, and Shih (2014,) Charness, Cobo-Reyes, and Jiménez (2014), Adnan, Arin, Charness, Lacomba, and Lagos (2022).

[5] See also Charness and Dufwenberg (2010) and Di Bartolomeo, Dufwenberg, and Papa (2019b).

when communication was permitted were dramatically better than without communication. Furthermore, direct statements of intent (promises) were particularly useful. If people interacting with out-group members are less favorable or trusting than with in-group members (which we do find in our data) might it be possible to transform people seen as out-group members to people seen as in-group members, with concomitant gains to the group or society?

We build on the papers focusing on the minimal-group paradigm. In a typical minimal-group experiment, subjects are randomly assigned to groups, which are intended to be as meaningless as possible; one such assignment used reflected whether people preferred paintings by Klee or by Kandinsky. Many experiments confirm and extend the Tajfel *et al*. (1973) finding that group membership creates in-group enhancement in ways that favor the in-group at the expense of the out-group. Subjects in these experiments – examples include Charness, Rigotti, and Rustichini (2007) and Chen and Li (2009) – show that group membership significantly affects individual behavior in treatments where groups are salient.[6] People tend to make choices that favor other people who are identified as in-group members at the expense of out-group members; again, this is true even with rather weak identity conditions.

We utilize a random-dictator game augmented by group membership, where membership is induced with the mechanism of randomly assigning each participant to a blue group or a red group.[7] Using an idea from Vanberg (2008),[8] we test the effects of communication by exogenous variations: we study how subjects who communicated directly would have behaved in a context in which they did not have direct communication. We do so by building a counterfactual where we look at how subject would behave in the shoes of others, but without personal involvement.

We test for differences in out-group allocations versus in-group allocations when there is direct communication and compare these differences to those present without direct communication. If there is an effect, in principle it could go in either direction. On the one hand, communication could increase the salient identity for people belonging to the same group or the between people belonging to different groups (for example, imagine there are groups formed of partisans for one of two sports teams and the discussion centers on this sport). On the other

[6] See also: Sutter (2009), Hargreaves Heap and Zizzo (2009), Chen and Li (2009), Kranton and Sanders (2017), Kranton, Pease, Sanders, and Huettel (2018), and Ciccarone, Di Bartolomeo, and Papa, 2020.
[7] In Charness and Villeval (2009), subjects chose yellow or green and formed groups based on this color choice.
[8] See also Ockenfels and Werner (2014) and Ciccarone, Di Bartolomeo and Papa (2020).

hand, direct communication may reduce the social distance (Charness, Haruvy, and Sonsino, 2003) between people belonging to different groups and reduce any in-group bias in the allocations made.[9] While one would hope (and hypothesize) that the latter tendency prevails, this is not at all a certainty.

Promises lead to more willingness to sacrifice money for both the pairs that remained intact and those that did not. As expected, one is more likely to make a favorable choice when paired with an in-group member. However, this gap vanishes entirely when both parties have pledged to choose the cooperative action; direct communication seems to bridge the gap between in-group and out-group members (partners and strangers). Perhaps this finding would apply as well in field environments. We also find interesting results regarding the motivation for honoring promises, with evidence for both guilt aversion and for a form of moral commitment. People are more willing to honor a promise that they have made than one that was made by another party in the same role, although there is some evidence that behavior is affected even in the latter case. Perhaps one feels guilt from not keeping the promise made by a peer. A particularly interesting result is that motivations differ between in-group and out-group pairs and that this difference drives an observed reduction in the in-group favoritism.

The remainder of this article is organized as follows. We present the experimental design and procedures in section 2, along with two models of social identity. We discuss the messages sent and received in section 3. Our experimental results are shown in section 4. We provide a discussion and conclude in section 5.

2. Experimental hypotheses, design, and procedures

2.1 Hypotheses

The first model of social identity in economics was developed by Akerlof and Kranton (2000), who propose a neoclassical utility function with identity associated with different social categories, as well as a prescription for behavior. Deviations from the prescription lead to disutility. Prescriptions indicate the behavior appropriate for people in different social categories in different situations. Their versatile framework has been applied to analyses of gender discrimination, the economics of poverty and social exclusion, the household division of

[9] Ciccarone, Di Bartolomeo and Papa (2020) found in-group favoritism in the same context without communication.

labor (Akerlof and Kranton 2000), the economics of education (Akerlof and Kranton, 2002), and the economics of organization (Akerlof and Kranton, 2005).

To formally endogenize the choice of group identities and norms, Shayo (2009) analyzes individuals' decisions to identify with social groups, focusing on the effects of social status and social distance. He presents (1) as a utility function for an individual *i* with group *J* that depends on agents' actions (a):

$$U_{i,J}(a) = \pi_i(a) - \beta_j d_{iJ}(a) + \gamma_i S_J(a), \quad (1)$$

where π_i is *i*'s material payoffs, d_{iJ} is *i*'s *perceived distance* from group J, and S_J is the *status* of group J. The parameters, β_i and γ_i are positive weights on social distance and group status, respectively, which can vary across individuals. Each agent is characterized by a vector of attributes $q_i = (q_i^1, q_i^2, \cdots, q_i^H)$. A social group *J* is characterized by the attributes of a *prototype* member, denoted q_J, i.e., an individual holding the average group attributes. The status of group *J* is determined through social comparison with other groups along valued dimensions of comparisons. Suppose Π_J and Π are measures of group *J* and its reference group's material payoffs, respectively. The status of group *J* can be characterized as $S_J = S(\Pi_J, \Pi, \sigma_J)$, where σ_J captures other determinants of group *J*'s status.

We see two dimensions in which the Shayo (2009) model applies to our setting. First, the perceived distance is likely to be affected by the experience of promises having been made. If the perceived distance between in-group members is small in the first place, it is reasonable to expect a larger effect from promises in the case of out-group pairs, so that the perceived distance d_{iJ} in (1) is larger for out-group pairs than for in-group pairs. Second, a social group *J* in which promises are featured may well have higher status S_J in (1) in the eyes of an individual, thereby increasing the utility of belonging to group *J*.

We consider seven hypotheses.

H1: *In-group bias*. Without direct communication, the average *Roll* rate of dictators matched with in-group recipients (color match) will be higher than the *Roll* rate of those matched with out-group recipients (color non-match).

Based on Ciccarone, Di Bartolomeo and Papa (2020), who found in-group favoritism in the same context without communication, we expect H1 to hold. This also follows from there being less social distance for in-group members due to their common identity.

H2: *Effect of chat*. We predict that staying matched with the person with whom one had become acquainted by direct chat will lead to a higher *Roll* rate for dictators, both with color matches and color non-matches (in-group and out-group).

In other words, H2 tests a straight communication effect, separate from identity considerations. However, combining H1 and H2 to evaluate the effect of chat on in-group bias involves some nuance. In our design, communication is a treatment that transforms strangers into fellows. The effect of the treatment on in-group bias is captured by a difference-in-difference comparison. We predict that mutual promises reduce the perceived distance between in-group and out-group members to something closer to that with two in-group members. Again, it is also possible that one considers that a group where mutual promises are made might have more status S_J, strengthening the effect. More formally:

H3: *Effect of chat on in-group bias*. Defining in-group bias as the difference between the average *Roll* rate of dictators in good matches and that of dictators in bad matches, we predict that in-group bias for fellow pairs will be less than that for stranger pairs.

Concerning H1-H3, if fellows were to exhibit little or no in-group bias and communication were to increase the *Roll* rate, the effects of the chats on the *Roll* rates should then be greater for out-group pairs then for in-group people if direct communication transforms strangers into fellows. Our inclination is to presume chat will "transform strangers into partners", thereby reducing any difference found without chat. Again, in principle this could go the other way, since people might instead comment on group differences, thereby exacerbating them.

We can investigate the mere effects of communication by comparing fellows and strangers. However, communication can be further qualified by inspecting its contents. In this respect, we can expect that promises could have specific effects on in-group bias. Clearly, classifying the content of communication involves a certain degree of subjectivity, making the

investigation more challenging. The next section presents the classification strategy. Here we just describe the hypotheses that we consider.

We begin with promises, by which we mean that both parties assure each other that they would *Roll* if they were to become the dictator.[10] We presume that the effects of the chat are stronger when we only restrict our attention to the promises. This prediction follows from the notion that these direct promises lead to a *per se* reduction in perceived social distance d_{iJ} (and a possible increase for the in-group's status) for out-group members, but little or no effect on perceptions of in-group members.

We again note that the effects of the promises on the *Roll* rates should be greater for out-group pairs than for in-group people if we are to observe decreased in-group bias. The idea is that the perceived social distance between the parties is already small for in-group pairs, but it is much more substantial for out-group pairs. Having made mutual favorable assurances reduces this difference in the latter case. Formally, we test the following:

H4: *Effect of promises on in-group bias*. We speculate that in-group bias will decrease significantly when promises are initially made.

Since our discussion is centered on promise keeping, we also look at its motivations. Two potential explanations can be stressed : i) moral commitment, promisors only keep their own word independently of their second-order beliefs, i.e., other people's expectations (Vanberg, 2008); ii) expectation-based explanation, promises can fuel second-order beliefs and promise keeping is then the resulting equilibrium of a psychological game where subjects are guilt averse (Charness and Dufwenberg, 2006).[11] Vanberg (2008) proposes a simple test for moral commitment, which is based on the same information structure used in our paper. He compares the behavior of dictators who consider whether to keep their own promises when there is no switch to that dictators who consider whether to keep promises involving another party. Here and there, the switch and asymmetric information imply that both these kinds of dictators should have the same second-order beliefs, everything else being equal. Therefore, evidence for moral

[10] From now on, we use "they" rather than "he or she".

[11] Note that both motivations are consistent with an observed correlation between promise keeping and second-order beliefs.

commitment is provided when dictators only keep their own promises. He finds evidence in favor of moral commitment.

We use Vanberg's test to explore motivation for promise keeping within in-group (or out-group) dictators. In keeping with Vanberg (2008), although in a different context, we suspect that moral commitment affects dictator behavior. The hypotheses that follow are not generated by the social-identity models although they are consistent with them.

H5: *Promise-keeping motivations*. Moral commitment plays a significant role in promisors keep their promises.

Of course, communication dynamics can be more complex than the simple dichotomy that promise/no promise reveals. Hence, we also consider the effects of communication from a broad perspective, i.e., how *Roll* rates differ across different sub-categories that do not involve initial promises. One might intuitively expect that one attempt to create an agreement would be better than none, or:

H6: *Effect of one attempt to form an agreement versus none*. We predict that the *Roll* rate will be higher when one of the subjects in the chat attempted to formulate an agreement than when no one does.

We also expect that it matters who has attempted to make an agreement, leading to:

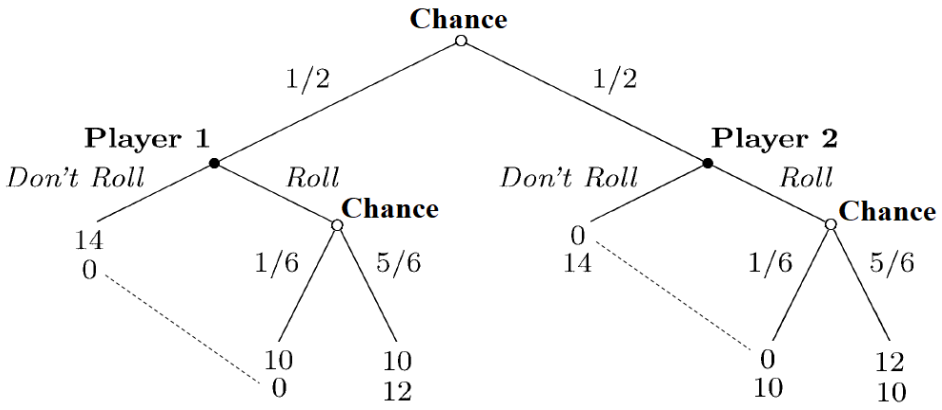
H7: *Effect of who attempts to create an agreement*. When there is only one party who has tried to create an agreement, we predict that the *Roll* rate will be higher when it is the dictator who has tried than when it is the recipient who has tried.

However, we note that using more finely-grained categories for communication content necessarily reduces statistical power when comparing message sub-classifications. Thus, as we shall see in this respect, some of our results should be seen as suggestive rather than conclusive.

2.2 Experimental design

Our experiment makes use of the binary-choice random-dictator game described in Figure 1.[12] It is a simple dictator game between two players, where it is initially unknown who will become the dictator but that each person has the same probability of this. We also assume that each recipient observes their own payoff, but not the paired dictator's action. The dotted lines show that the recipient cannot tell if a payoff of zero was the result of a selfish choice by the other player or the result of bad luck.

Figure 1. A binary-choice random-dictator game



We introduce four critical changes relative to the game in Figure 1. They are described below. The first two occur before the game is played, while the last two occur after *Nature's* choice.[13]

1. **Group membership (minimal group paradigm).** Consider a set composed of an even number *N* of subjects. At the beginning, each of the *N* subject is randomly assigned to one of two groups (Red or Blue). Then, *N*/2 pairs of subjects are formed randomly, and all subjects observe the color of their partner, so they know whether they belong to the same group ("color match") or not ("color non-match").
2. **Chat.** Each person is matched with another person in a pair. Within each pair, subjects are given the opportunity to communicate by chat. After any such communication take place, *Nature* chooses the paired-subject role of either dictator or recipient.
3. **Partner switch.** After subjects are told their roles, half of the recipients are randomly re-matched with a new dictator. We refer to the switched agents as *strangers* as they play

[12] We use a dictator game as in Güth *et al*. (2009), Ockenfels and Werner (2014), and Ciccarone. Di Bartolomeo and Papa (2020).
[13] See the next subsection for details

with someone who did not chat with them and refer to the non-switched agents as *fellows* since they play with someone with whom they have directly communicated. Fellows are thus acquainted through their earlier chat while strangers are not.

4. **Asymmetric information.** Recipients do not know whether they or not they have been switched, but before a dictator makes a choice, the dictator is informed if the initial paired person was switched. They can then: **a)** observe the color of the new recipient and the color of the new recipient's old partner, and **b)** read the communication that occurred between his/her new recipient and the person with whom he or she was initially matched. In this way, the dictator's beliefs about the beliefs of the recipient (second-order beliefs) are independent of the switching.

Finally, we use payoffs from Charness and Dufwenberg (2006). Each dictator (whether involved or not in a switch) chooses between *Roll* and *Don't Roll*, as in Figure 1. Choosing *Don't Roll* leads to the dictator receiving 14 tokens, and the recipient receiving nothing; choosing *Roll* leads to the dictator receiving 10 tokens, whereas the recipient receives 12 tokens with probability 5/6 and nothing with probability 1/6.

In a nutshell, one is initially matched with someone having the same or a different color and these counterparts could communicate with each other. However, all know that they could then play with someone else. At the end, each dictator could be paired with someone who exchanged a message with her (in the case of no-switch) or not (in the case of switch), i.e., we observe dictators in same-color or other-color matches, who play with someone who communicated with them (a fellow) or not (a stranger). Everyone had been told that the dictator will read the communication of a switched person if a switch has occurred. It is worth noting that we created this information structure to generate exogenous variation in communication. This occurs because the dictators know that recipients do not know whether they or not they are switched.[14] Hence, dictators' second-order beliefs should be independent of whether a switch occurred: all else equal, fellows and strangers should have the same beliefs.[15]

[14] Exogenous variations also occur, since switched dictators have access to the information about their new partners.

[15] The relevance of accounting for belief effects in dictator games is stressed by, among others, Khalmetski, Ockenfels, and Werner (2015), who point out how correlation between transfers and expectations can be either positive and negative, obscuring the effect in the aggregate.

Our interest is twofold. First, we are interested in studying the effect that direct communication has on favoritism relative to indirect communication. So, we compare the behavior of fellows to counterfactuals built on strangers. We also use a difference-in-difference between the in-group bias within strangers to that within fellows. Second, we investigate the effects of communication on in-group bias. We consider the relative effects of promises involving people who either belong or don't belong to one's own group. Note that the first issue does not require a communication classification, while the second issue does. Classification reflects promises to act favorably, i.e., when both the sent message and the received message convey an intent to choose *Roll*. Recall that people do not know their roles when chatting.

2.3 Procedures

The experiment was conducted at the Sapienza University of Rome (CIMEO Lab). The design involved 384 students (12 sessions, 10 rounds each, 32 subjects each), recruited using an online system. Upon arrival, subjects were randomly assigned to 32 isolated computer terminals.[16] Two assistants handed out instructions and checked that participants correctly followed the procedures.

The experiment design consisted of three stages: 1) group assignment (pre-session), where pairs of subjects observed their colors, 2) experimental session, and 3) final payment (post session). After the group assignment, subjects played the experimental session, which consisted of ten rounds, with perfect stranger matching. Each round implemented the following sequence of five stages:

1. *Communication*. Subjects were randomly matched to form 16 chatting pairs, with random determination of who would start the chat. Each chat consisted of four one-way messages in sequence. Each message could be of at most 90 characters.
2. *Role assignment.* At the beginning of the round, roles (dictator or recipient) were randomly assigned for each pair and subjects were so informed.
3. *Switching*. Half of the pairs were switched. Only dictators were informed of whether a switch occurred. Each switched dictator with a new recipient learned: a) the color of the new recipient, b) the color of the previous paired dictator of the new recipient, and c) the prior conversation that the new recipient had with the previous paired dictator.

[16] The experiment was designed using z-Tree (Fischbacher, 2007).

4. *Belief elicitation*. This stage has two parts: a) first-order beliefs: each recipient was asked to guess about if the paired dictator (after the switch) would choose to *roll*, and b) second-order beliefs: dictators when there was a switch were asked to guess the guess of the current recipient, for both those who were switched and those who were not.[17]
5. *Dictator's action*. The dictator made the choice of *Roll* or *Don't Roll*. All subjects were informed of their payoff for the round. A recipient could not be certain of the dictator's choice when the received payoff was zero.[18]

At the end of the session, subjects were paid. All subjects received a fixed show-up fee of 2.50 tokens. One of the rounds was randomly chosen for payment. The payoffs shown in Figure 1 were computed in tokens (where 1 token = 0.5 euro). Subjects were told truthfully that incentives for beliefs elicitation were provided for all rounds, except the one chosen for payment (removing an incentive to hedge by mis-reporting beliefs).[19]

[17] Belief elicitation is described in detail in the next section. We stress that we elicit beliefs after dictators know whether they have been switched or not, but before dictators make their choice. Hence, we elicit the beliefs of switched and non-switched dictators regarding the beliefs of their recipients, as in Di Bartolomeo *et al*. (2019b, 2020) and Vanberg (2008). Since switched dictators should infer the beliefs of the new recipients, elicitation is done after they read the chat of their partners. Again, dictators know that recipients do not know if they were switched.

[18] Recipients could obtain a zero payoff in two cases: (i) their dictator had chosen *Don't Roll*; (ii) their dictator had chosen *Roll* and the outcome of the die-roll was "1."

[19] A reader mentioned that we could have instead studied how concerns related to group identity effects the formation of trust and cooperation. Indeed, this is a key topic. However, our experimental design is not really geared to this question, so we suggest this would be a very useful and related additional paper.

3. Messages

The communication that took place between the participants is the heart of our experiment. Research assistants blind to our hypotheses catalogued all the messages.[20] In this study, it is worth remembering that communication is bilateral and has a back-and-forth nature. Moreover, participants chat before knowing their role, knowing that one of each pair will be a dictator and the other will be a recipient.[21]

We note that this protocol differs somewhat from those used previously. Charness and Dufwenberg (2006) used free-form chat with a one-page limit and only one message sent from the second mover to the first mover. We might expect more effective communication with more communication rounds, but that having a 90-character restriction on each message might limit this. In our back-and-forth communication protocol, participants may be inclined and able to strike a deal of conditional cooperation: "I'll pledge to choose *Roll*, if you also pledge to choose *Roll*." When both players do so, we find it natural to define a "promise" as a promise to keep this form of agreement.[22]

Operationally, we asked to the research assistants to classify each subject's message according to whether it conveyed a pledge.[23] Each subject is considered a promisor if and only if a) their message conveys such a pledge; and b) the message sent by their counterpart also contains a pledge stating that they would choose *Roll*. In other words, we observe a "promise" (*P*) if and only if both sides state their intention to roll (*RR*), where *R* refers to the pledge. By contrast, we classify as non-promises (*NP*) all other communication outcomes, i.e., when only one (or no) party discloses such an intention (*RN*, *NR*, *NN*).[24]

Clearly our assumption is strong, since the communication dynamics may also affect behaviors. A dictator can infer some information about the kind of recipient she faces from

[20] The research assistant was not involved in the design and execution of the experiment. Indeed, we asked three research assistants to classify messages and *ex ante* randomly choose the classification of one of them for the experiment. The different classifications were however strongly correlated (with a high Cronbach alpha value of 0.8880.)

[21] We note that this communication protocol differs somewhat from those used previously. First, we know of no experiment featuring chat before one's role has been determined. Second, Charness and Dufwenberg (2006) used free-form chat with a one-page limit and only one message sent from the second mover to the first mover.

[22] This random-dictator design, adapted from Vanberg (2008), is therefore not the same as Charness and Dufwenberg (2006). On this point, see Di Bartolomeo *et al*. (2021), where unilateral one-shot communication is instead considered.

[23] As in Vanberg's protocol, each pair of messages sent by the same subject in a round was treated as a unit.

[24] All the examples are from the experiment chats (translated from Italian.)

reading his communication. In principle, that information can be relevant. For instance, she could be more likely to support a recipient who is claiming to be ready to roll. For the sake of robustness, we will look at communication dynamics in Section 4.6, where we focus on H6 and H7. However, it is not possible to control for everything; concerns such as rounds, who started the chats, its duration, and so on could be relevant as well.[25]

Some examples may help to clarify. We present these in order of the intuitive likelihood the dictator will subsequently choose to roll. Either player could become the dictator at this point. Examples 1a and 1b involve a promise (*P*), all the other examples are no-promises (*NP*).

Example 1a. Both people state that they would like to roll. (*RR*)

Player No 1 says "If I am Player A [dictator], I will roll the dice."
Player No 2 then replies, "I would roll the dice, too."
Player No 1 says "So, let's roll, Player A [dictator] will roll!"
Player No 2 says "OK!"

Example 1b. (*RR*)

Player No 1 says "do you pledge to roll the die?"
Player No 2 then replies, "Yes, you too, though."
Player No 1 says "I pledge!"
Player No 2 says "OK!"

Example 2a. The drawn dictator pledges to roll while the drawn recipient did not. (*RN*)

Player No 1 (future dictator) says "I will roll!"
Player No 2 replies "I am not sure."

Example 2b. (*RN*)

Player No 2 says: "New washing machine launched on the market: 12 dead, 4 injured. Will you roll the die?"
Player No 1 (future dictator) replies "I always roll the die."
Player No 2 says "You like to live on the edge, huh?"
Player No 1 replies "Maybe."

Example 3a. The drawn dictator made no cooperative statement, but the drawn recipient did (*NR*)

Player No 2 says "I will roll!"
Player No 1 (future dictator) replies "I am not sure."

[25] Controlling for all taxonomies reduces the power of the tests, since the number of observations in each resulting case drops correspondingly.

Example 3b. (*NR*)

Player No 2 says “I roll the die.”
Player No 1 (future dictator) replies “I will not.”
Player No 2 says “alright!!!”
Player No 1 replies “ok”

Example 4a. (*NN*)

Player No 1 says “Life is an abyss.”
Player No 2 replies “Yes but that’s how life goes, I’m a constant carpe diem; so, I don’t waste time and I live.”
Player No 1 says “Very well”
Player No 2 replies “While you observe the abyss, the abyss observes you; if you roll the dice, I’ll give you a rope, so you don’t fall.”

Example 4b. Neither person disclose a positive intent. (*NN*)

Player No 1 says “I will not roll the dice.”
Player No 2 replies “Me neither.”

Our experiment involves 384 subjects playing 10 rounds, that is, 1920 chats. Therefore, our sample consists of 3,840 messages. Among the chats, 2,226 (58%) were considered promises. It is worth noting that messages considered here include both dictators and recipients, since chats occurred before subjects know their rule. With in-group matches, 1,210 of 2,016 chats (60.0%) were promises; with out-group sample, 1,016 out of 1,824 (55.7%) were promises.

Table 1 shows the details, i.e., promise rates made with out-groups and in-groups in each session. The rate of promises was slightly higher for in-group matches, but not significantly so ($Z = 1.18$, $p = 0.239$). We see no significant in-group bias for promises made.

Table 1. Matching and promises by session (1920 obs.)

	Session												
Category	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**All**
Out-group	.62	.54	.54	.58	.78	.61	.49	.59	.21	.47	.67	.59	**.56**
In-group	.58	.60	.67	.67	.77	.73	.54	.71	.45	.38	.52	.58	**.60**

4. Results

We first present our results on belief elicitation in the first two sub-sections. The final two sub-sections discuss in detail the actions chosen by the participants.

4.1 Elicitation of first- and second-order beliefs

After communicating, recipients were asked to guess what their (unknown) dictators would choose to do. They had been told the switching probability is always 50% in each treatment. Thus, they were aware that their paired subject could be switched according to that probability. Recipients could make their guess by ticking one of the five-point scale in Table 2. This scale is the same as in Vanberg. Beliefs are then re-scaled to 1.00, 0.75, 0.50, 0.25, and 0.00. Thus, the numbers shown in Table 3 below represent the averages of the dictators' re-scaled responses.[26]

Table 2 – Incentives for first-order belief elicitation

The dictator will	choose *Roll*			choose *Don't Roll*	
	Certainly	Probably	Unsure	Probably	Certainly
Please tick your guess	◯	◯	◯	◯	◯
Your earnings if the dictator					
chooses *Roll*	0.65 tokens	0.60 tokens	0.50 tokens	0.35 tokens	0.15 tokens
chooses *Don't Roll*	0.15 tokens	0.35 tokens	0.50 tokens	0.60 tokens	0.65 tokens

After dictators were told whether their paired recipient (in-group or out group) had been switched or not, they were asked to guess his guess. Specifically, they had to guess which of the five points of Table 2 had been ticked by their counterpart. Correct guesses earned 0.50 tokens.

4.2 Second-order beliefs and exogenous variations

Our design involves several exogenous variations, which we shall discuss. First, we wish to verify that the data are consistent with these variations. We test this with the elicited second-order beliefs.

[26] The payoffs correspond to a quadratic scoring rule for probabilities 85%, 68%, 50%, 32%, and 15%; note that risk neutrality implies that quadratic scoring yields flat payoffs as probabilities approach one (see Vanberg, 2008: 1472). We also verify the robustness of our results to the quadratic scoring rule. Results are available upon request.

Table 3. Matching, second-order beliefs (1920 obs.)

	Session												
Category	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**All**
(1) Strangers, out-group	.70	.62	.75	.60	.74	.64	.63	.64	.68	.78	.74	.65	**.68**
(2) Fellows, out-group	.64	.57	.68	.61	.72	.68	.65	.73	.61	.77	.84	.62	**.68**
(3) Strangers, in-group	.67	.64	.71	.63	.71	.49	.65	.79	.71	.72	.70	.65	**.67**
(4) Fellows, in-group	.63	.67	.72	.69	.72	.64	.61	.78	.82	.73	.75	.68	**.70**

Table 3 shows average second-order beliefs of out-group and in-group dictators by session,[27] distinguishing in both cases between strangers and fellows. We conduct our non-parametric analysis using Wilcoxon signed-ranks tests on the session-level data.[28] We also conservatively report two-tailed tests. Table 3 shows that, all else equal, there are no differences in the second-order beliefs of stranger and fellow dictators for both the cases of out-group and in-group; one would hope that this is true, since dictators know that the recipients don't know if they had been switched. We find $Z = 0.39$, $p = 0.695$ across out-groups and $Z = -1.69$, $p = 0.091$ across in-groups. Table 3 also shows two other exogenous variations in group membership. First. there are no differences in the second-order beliefs of out-group dictators and in-group dictators with strangers. Our test finds $Z = -0.94$, $p = 0.347$. Second, there are no differences with fellows, since this test gives $Z = 0.98$, $p = 0.327$.

Table 4 verifies two further exogenous variations in group membership. The first four rows show second-order beliefs of dictator fellows who either participated (*P*) or did not participate (*NP*) in a promise, distinguishing between the out-group and the in-group cases. In the two last rows, the table also reports the second-order beliefs of dictator strangers who made participated in a promise are re-matched with someone who did not receive one (*NPR*). Regarding fellows who participated in a promise, there are no differences in the second-order beliefs of out-group (51%) and in-group dictators (55%), $Z = -0.63$, $p = 0.530$. Similarly, among dictators who did not participate in making a promise, there are no differences in the second-order beliefs of out-group (81%) and in-group dictators (80%), $Z = 0.55$, $p =$

[27] Recall that we are looking for exogenous variations in being acquainted in both out-group and in-group cases.
[28] Throughout this paper, all tests are two-tailed Wilcoxon signed-ranks tests on session-level data, unless otherwise specified. All *p*-values are rounded to the nearest three decimal places.

0.583. Strangers (the last two rows of Table 4) show an exogenous variation in group membership of fellows who participated in making a promise who made a promise and are re-matched with someone who was not involved in a promise. There are no differences in the second-order beliefs of out-group (53%) and in-group dictators (56%): Z = -1.02, p = 0.308.

Table 4. Average second-order beliefs of selected dictators' categories by session (1169 obs.)

	Session												
Category	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**All**
(1) Out-group, fellow, *NP*	.30	.38	.61	.45	.56	.31	.53	.56	.54	.57	.68	.50	**.51**
(2) In-group, fellow, *NP*	.49	.43	.53	.63	.43	.32	.50	.52	.77	.60	.62	.54	**.55**
(3) Out-group, fellow, *P*	.86	.71	.77	.69	.77	.87	.91	.89	.80	.96	.92	.70	**.81**
(4) In-group, fellow, *P*	.72	.83	.82	.73	.81	.74	.71	.88	.92	.86	.85	.78	**.80**
(5) Out-group, stranger, *NPR*	.68	.55	.61	.47	.75	.29	.40	.38	.58	.88	.72	.32	**.53**
(6) In-group, stranger, *NPR*	.59	.63	.75	.56	.79	.35	.56	.45	.61	.54	.58	.54	**.56**

Finally, the first four rows of Table 4 show that promisors who had experienced promises have higher beliefs compared to the others (0.80 vs. 0.55 for the in-group sample, Z = 3.06, p = 0.002; 0.81 vs. 0.51 for the out-group sample: Z = 3.06, p = 0.002), consistent with the notion that guilt aversion fosters promise keeping (Charness and Dufwenberg, 2006). Moreover, as one might expect, the beliefs of dictators who did not experience promises and those of dictators who experienced promises but are rematched who someone who was not involved in a promise do not differ at a statistically-significant level (0.56 vs. 0.55 for the in-group sample, Z = 0.75, p = 0.456; 0.53 vs. 0.51 for the out-group sample: Z = 1.02, p = 0.308).

Tables 3 and 4 together confirm that experimental data are consistent with our (theoretically-designed) exogenous variations in being acquainted and in group membership.

4.3 Direct communication and social identity

Table 5 reports the roll rates for dictators by whether there was communication and whether there was a switch. Recall that we us two-tailed Wilcoxon signed-ranks tests on the session-level data. Again, our design involves exogenous variation in communication. The comparison between row (1) and row (2) provides evidence in favor of in-group favoritism within strangers, as expected. The average *Roll* rate of dictators in good matches (50%) is

significantly higher than that of dictators in bad matches (40%), with Z = 2.94, $p = 0.003$. The comparison between row (3) and row (4) also provides evidence in favor of in-group favoritism with dictator fellows.[29] The 58% average *Roll* rate of dictators in good matches is statistically higher than the 50% average *Roll* rate of dictators in bad matches ($Z = 2.04$, $p = 0.041$). Summing up, favoritism is observed among both strangers and fellows. Dictators in color matches are more likely to roll than are dictators in non-color matches. Overall, we find considerable support for in-group favoritism (H1 and H2).

Table 5. *Roll* rates by session and category (1920 obs.)

	Session												
Category	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**All**
(1) Out-group, strangers	.46	.33	.31	.18	.46	.31	.28	.33	.56	.59	.59	.33	**.40**
(2) In-group, strangers	.54	.44	.41	.39	.59	.27	.37	.54	.66	.73	.63	.44	**.50**
(3) Out-group, fellows	.57	.49	.43	.46	.54	.34	.31	.54	.57	.54	.71	.54	**.50**
(4) In-group, fellows	.67	.56	.47	.31	.53	.53	.42	.64	.82	.71	.69	.56	**.58**

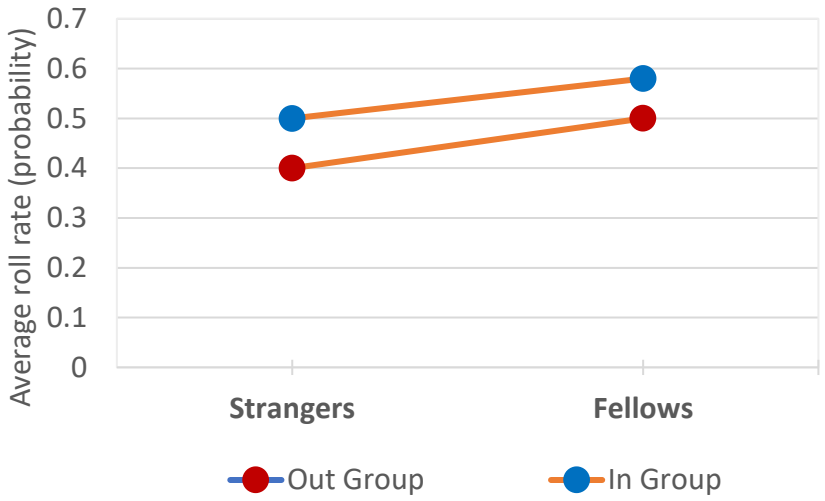
Does having had direct communication affect the behavior of dictators? For this, we focus on rows and test H2. Staying matched and having therefore had direct communication increases the dictators' average *Roll rates* in both color matches and non-matches. Comparing strangers to fellows, with color non-matches (out-group), the average roll rate increases from 40% to 50%; with color matches (in-group), this rate increases from 50% to 58%. The differences are easily significant with these conservative tests ($Z = 2.75$, $p = 0.006$ and $Z = 2.35$, $p = 0.019$, respectively). Comparing across stranger categories shows that the *Roll* rate was higher for in-groups in 11 of 12 sessions, while this rate was higher for in-groups in 9 of 12 sessions across fellow categories. Thus, our results strongly support H2. Direct communication leads to more people choosing *Roll*, with similar effect sizes for color matches and non-matches.

Finally, we test the effect of having become acquainted through chat on the in-group bias (H3), i.e., we test a different effect of chat conditional to the kind of match occurred. Our test is shown in Figure 2. The average *Roll* rate for out-group dictators increases from 0.40 to 0.50 with

[29] Recall that in both cases, average second-order beliefs are not significantly different at the row level due to our exogenous variation, (cf. Section 4: Table ~~3~~).

the treatment (acquainting), while it increases from 0.50 to 0.58 for in-group dictators. There is no significant difference between these differences (0.10 vs. 0.08: $Z = 0.94$, $p = 0.347$). Thus, our results do not provide support for the notion that in-group bias leads to different effects for direct pre-play communication between agents.

Figure 2 – Dictators average roll rates



4.4 Promises and social identity

We now complement the result of previous section by looking at the content of the communication, which was described in Section 3. Table 6 reports the average *Roll* rates of fellows who were involved in an initial promise (*P*) or did not (*NP*) and those of strangers who were involved in an initial promise and are re-matched with someone who did not receive one (*NPR*). The table also considers whether there was a color match.[30]

Looking at the table, a bird's eye view shows us that, among fellows who do not mutually declare intentions to act favorably, dictators are 11 percentage points more likely to sacrifice own payoffs when they are paired with same-color recipients than with different-color recipients (0.39 vs. 0.28: $Z = 2.20$, $p = 0.028$). Conversely, there is no gap when promises are considered

[30] Rows (1) and (2) refer to those who were not involved in a promise (*NP*) during pre-play communication [i.e., *RN*, *NR*, or *NN*], rows (3) and (4) to those who were. Recall that, because of our exogenous variations, average second-order beliefs of in-group and out-group these dictators are not significantly different within the two categories (promisors and no promisors). See Session 4: Table 4.

(0.69 vs. 0.68: $Z = 0.75$, $p = 0.456$). Comparing across no promisor categories shows that the *Roll* rate was higher for in-groups in 10 of 12 sessions (row (2) vs. (1)), while the roll rate was higher for in-groups in 7 of 12 sessions with promisor categories (row (4) vs. (3)).

Table 6. Average roll rates of selected dictators' categories by session (1169 obs.)

	Session												
Category	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**All**
(1) Out-group, fellow, *NP*	.36	.20	.26	.09	.38	.00	.21	.12	.56	.29	.45	.36	**.28**
(2) In-group, fellow, *NP*	.47	.28	.31	.22	.20	.18	.29	.08	.74	.55	.53	.37	**.39**
(3) Out-group, fellow, *P*	.71	.70	.63	.63	.59	.52	.55	.94	.60	.78	.83	.67	**.68**
(4) In-group, fellow, *P*	.79	.74	.55	.37	.63	.65	.54	.85	.94	.87	.81	.69	**.69**
(5) Out-group, stranger, *NPR*	.36	.40	.27	.25	.25	.14	.10	.20	.33	.75	.38	.43	**.30**
(6) In-group, stranger, *NPR*	.50	.25	.71	.58	.67	.31	.50	.40	.78	.50	.44	.54	**.50**

Do the data support the idea that promises eliminate the in-group bias? Indeed, people who made promises were much more likely to sacrifice than people who made no promises, both for color matches and non-matches. In Table 6, our tests give (0.68 vs. 0.28: $Z = 3.06$, $p = 0.002$; 0.69 vs. 0.39: $Z = 3.06$, $p = 0.002$) for comparisons between rows (1) and (3) and between rows (2) and (4), respectively. Summing up, we find an in-group bias among fellows without promises. However, there is no in-group bias amongst fellows who were involved in an initial promise.

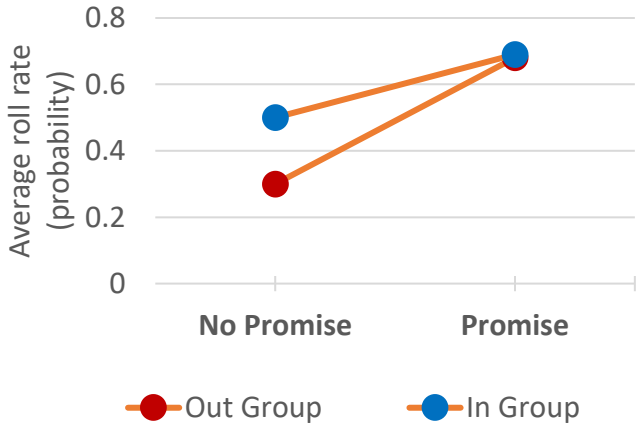
The previous comparison is between subjects who were involved in an initial promise and subjects who do not. Another comparison is to consider how the dictators who were involved in an initial promise would have behaved in the absence of it. This exercise, in our design, can be done by constructing a counterfactual. Specifically, we use the data stemming from our switching mechanism. The comparison is then between a kind of dictator who were directly involved in an initial promise (*fellows, P*) and how the *same* kind of dictator would have behaved in the absence of the direct promise (*stranger, NPR*). We construct such a counterfactual to obtain a fair comparison.[31]

[31] We also note that people who made a favorable declaration may have different attitudes than those people who did not. In-group promisors who read a non-promise are more likely to roll than people who did not promises (0.50 > 0.39: $Z = 2.43$, $p = 0.015$). However, no statistically significant difference is present in the out-group subsample (0.30 > 0.28: $Z = 0.86$, $p = 0.388$).

In this respect, Table 6 shows no group bias for dictators (*fellows*, *P*) who made a promise, as was previously mentioned; while the comparison between row (5) and row (6) indicates a substantial and significant in-group bias (0.50 vs. 0.30: $Z = 2.20$, $p = 0.028$) for counterfactual dictators (*strangers*, *NPR*); comparing across promisors who do not read promise categories shows that the *Roll* rate was higher for in-groups in 10 of 12 sessions. It seems that promises serve to effectively eliminate the in-group bias observed in the absence of direct communication and promises.

Our last result needs to be investigated further. It implies that promises (*P*) are more effective in the out-group people. The idea can be better captured by the difference-in-difference comparison shown in Figure 3. In the figure, the in-group bias is measured by the difference between the average roll rate of in-group and out-group dictators. In the figure we compare the in-group bias associated with dictators (*fellows*, *P*) who were involved in an initial promise and that associated to counterfactual dictators (*strangers*, *NPR*). The former is measured by the distance between the blue and red points on the promise line, while the latter is measured by the distance between the blue and red points on the no-promise line.

Figure 3 – The effects of making a promise on in-group favoritism

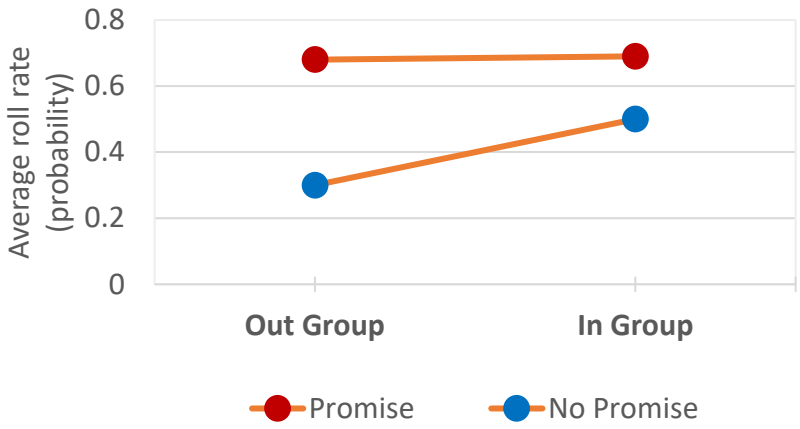


The difference-in-difference comparison provides evidence that promises are more effective for the out-group people (i.e., 0.20 vs. 0.01: $Z = 1.96$, $p = 0.049$), so, notwithstanding the fact that promises greatly foster *Roll* rates, they also eliminate the in-group bias. In the no - promise case, out-group dictators are less likely to roll than the in-group ones (although they have the same second-order beliefs). After being involved in a promise, all dictators are more

likely to *Roll* but a promise is more effective when the counterfactual situation was less favorable to the recipient (again, although second-order beliefs are the same).

In a similar vein, we can also test the impact of a promise in good and bad matches. Accounting for color matches, Figure 4 compares individuals who have made a promise to those who have made one but who find themselves in a counterfactual situation where no promise was made. This comparison again supports the idea that promises are more effective with out-groups than with in-groups (i.e., 0.68 - 0.30 = 0.38, while 0.69 - 0.50 = 0.19, with $Z = 1.96$, $p = 0.049$). Overall, we find support for H4 (i.e., promises decrease favoritism). Summing up, promises to *Roll* do not affect behavior independently from the in-group or out-group status –they are more effective in the latter and tend to reduce or even eliminate the in-group bias.

Figure 4 – The effects of membership on promise keeping



4.5 Exploring motivations

We now explore the rationale for promise keeping distinguishing motivations of outgroup dictators from those of ingroup dictators. In a nutshell, we apply Vanberg's test (H5) within the two subsamples to separately test their motivations. We focus on dictators involved with promises, comparing the behavior of non-switched dictators to the behavior of those dictators rematched with a recipient who has received an assurance by another person. To obtain a fair comparison in terms of second-order beliefs, we only account for the dictators re-matched with

recipients who were previously outgroup (ingroup) when the outgroup (ingroup) comparison is considered.[32]

Table 7. Average Roll rates of selected dictators with promises (718 obs.)

	Session												
Category	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**All**
Outgroup dictators													
(1) Non-Switched[(a)]	.71	.70	.63	.63	.59	.52	.55	.94	.60	.78	.83	.67	**.68**
(2) Switched[(b)]	.60	.67	.14	.00	.50	.60	.56	.71	na	.67	.70	.44	**.53**
Ingroup dictators													
(3) Non-switched[(c)]	.79	.74	.55	.37	.63	.65	.54	.85	.94	.87	.81	.69	**.69**
(4) Switched[(d)]	.73	.67	.60	.43	.55	.60	.67	.80	1.0	1.0	1.0	1.0	**.69**

Notes: (a) Non-switched outgroup dictators who participated in a promise; (b) Switched outgroup dictators who participated in a promise and were rematched with a recipient who was also participating in a promise with another outgroup partner. (d) Non-switched ingroup dictators who participated in a promise; (d) Switched ingroup dictators who participated in a promise and were rematched with a recipient who was also participating in a promise with another ingroup partner.

Roll rates are provided in Table 7. We note that within the ingroup (outgroup) sample, second-order beliefs are the same between switched and non-switched dictators, as we have argued that they should be.[33] It appears that ingroup and outgroup promisors have somewhat different motivations in keeping their promises. Our result provide evidence that the behavior of outgroup promisors is consistent with a moral commitment, i.e., they are significantly more likely to keep their own promises (68% vs. 53%: $Z = 2.58$, $p = 0.009$). H5 receives support for outgroup dictators. Conversely, the behavior of ingroup dictators does not show evidence of moral commitment: all else equal, they also keep promises made by others (69% vs. 69%: $Z = 1.09$, $p = 0.272$]. H5 is thus not supported for ingroup dictators. Their behavior is consistent with choices only being driven by a sense of guilt.

As shown in the previous sections, the reduction of the ingroup bias reflects promises of outgroup subjects being relatively more effective than those of ingroup subjects. So, our result on

[32] We verify the exogenous variation by testing the equality in the second-order beliefs. In-group promisors who read a promise by a new in group recipient have the same second-order beliefs than in group who made promises (0.80 vs. 0.80: $Z = 0.31$, $p = 0.754$). Out-group promisors who read a promise by a new out group recipient have the same second-order beliefs than out group who made promises (0.83 vs. 0.81: $Z = 0.18$, $p = 0.859$).

[33] SOBs are reported in the Appendix

motivation implies that the ingroup bias reduction is thus mainly driven by the moral-commitment effect that characterizes the motivations of outgroup dictators.

A difference-in-difference approach offers much the same result. We compare the 15 percentage-point increase in own promise-keeping in the outgroup sample to the 0% difference in the ingroup sample, finding a statistically-significant nonparametric difference-in-difference ($Z = 2.31$, $p = 0.021$]. So, the increase in promise keeping when one is asked to honor one's own pledge as part of a promise instead of a pledge made by another as part of a promise is greater for outgroup dictators than the increase for the ingroup ones.

We could argue that the sense of being in a group might serve as a substitute for a promises. For example, if you have initial x-group pairing and a promise, followed by a switch to an in-group new match, then predict that more is given than if you have initial x-group pairing and a promise, followed by a switch to an out-group new match.

4.6 Differences in behavior across non-promise sub-categories

We only include promises to keep agreements to roll in our classification. However, there are three sub-categories (*NN*, *NR*, *RN*) that comprise the non-promise *(NP)* classification. Our hypotheses H6 and H7 consider differences in rates across these sub-categories. One might expect the lowest rate to occur when no pledge is made (H6). With exactly one person declaring favorable intentions, a second intuition is that the *Roll* rate should be higher when the person who makes this declaration becomes the dictator (H7).

Table 8 shows the observed *Roll* rates.

Table 8. Average roll rates by non-promise (*NP*) categories by session (600 obs.)

	Session												
Category	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**All**
(1) Out-group, fellow, *NN*	.38	.00	.29	.25	.33	.00	.23	.00	.53	.29	.60	.43	**.31**
(2) In-group, fellow, *NN*	.36	.25	.20	.50	.00	.00	.33	.25	.67	.36	.46	.27	**.38**
(3) Out-group, fellow, *NR*	.00	.00	.00	-	.40	.00	.00	.00	.50	.00	1.00	.17	**.18**
(4) In-group, fellow, *NR*	.60	.25	.29	.00	.00	.20	.38	.00	1.00	1.00	.67	.00	**.29**
(5) Out-group, fellow, *RN*	.00	.50	.30	.00	-	.00	.30	.30	1.00	.50	.20	1.00	**.31**
(6) In-group, fellow, *RN*	1.00	.33	.50	.33	.67	.33	.14	.00	.86	.80	.67	.80	**.49**

Note that we have just over half the observations in Table 6. This necessarily reduces the resulting statistical power. Nevertheless, we do have suggestive evidence regarding hypotheses H6 and H7. First, the *Roll* rates with fellows are in fact *higher* when no pledges were made (*NN*) than when only the recipient has tried to find an agreement (*NR*); the overall comparisons are 0.18 versus 0.31 for out-group matches and 0.29 versus 0.38 for in-group matches. However, while the differences in percentage points are about the same as the significant percentage-point differences shown in Figure 2, these differences are not significant with our conservative session-level tests and limited data.[34] The signed-rank test gives $Z = -1.26$ and $p = 0.207$, $Z = -0.75$ and $p = 0.454$ for out-group and in-group matches, respectively. For in-group versus out-group fellow recipients who unilaterally pledge to roll, we have 29% vs. 18%, $Z = 1.47$ and $p = 0.142$; for in-group versus out-group fellows who did not make pledges, we have 38% vs. 31%, $Z = 0.43$ and $p = 0.665$. These differences are about the same size as those without communication, so we find no support for the reduction of in-group bias in these sub-categories. Overall, there seems to be a *decrease* in favorable actions for pairs with one favorable declaration rather than zero, although there is insufficient power for statistical significance.

How might we explain this result, which contradicts our original intuition? Data from Charness (2000b) may shed some light on this. In that experiment, subjects bargained over a number (price) and the total pie shrank with delay in reaching an agreement (or no agreement). Prior to this, subjects were categorized as either low or high types based on allocations made in a dictator game, with high types more generous, and were informed of their own type and that of their counterpart. While the cost of dispute (the shrinkage of the pie) was lowest when two high types were paired, this was highest with cross-pairings rather than with a match of two selfish types. People appear to take the action in the dictator game as a proxy for one's type and feel less willing to compromise with a person of the other type. We suspect that this is a common reaction.

Second, we consider *Roll* rates when precisely one person in a pair has attempted to find an agreement (*RN* and *NR*). The rate is substantially higher when the dictator rather than the recipient has made the favorable declaration, 0.18 versus 0.31 for fellow out-group matches, $Z =$

[34] Less stringent statistical tests (e.g., considering each person to be one independent observation) do confer statistical significance in most cases. In addition, many of our tests (those with $Z > 1.39$ for two-tailed tests and $Z > 1.16$ for one-tailed tests) would be significant with the same patterns in the data but twice the sample size.

1.67 and $p = 0.095$); this comparison is 0.29 versus 0.49 for fellow in-group matches, $Z = 1.46$ and $p = 0.145$). One feels more bound by one's own declaration (*RN*) than by a declaration made by another party (*NR*). In fact, the difference across these declarations widens with in-group matches, from 0.13 to 0.20; it seems that one is less likely to violate one's assurance when it is made to an in-group member (0.20 versus 0.13, $Z = 1.24$ and $p = 0.214$.)

6. Conclusion

One's social identity has become a major feature in our contemporary society. In recent years, identity has been used more and more as a wedge to separate subgroups. While identity is indeed often seen as a divisive force, it may potentially instead be utilized to benefit society (see Charness and Chen, 2020, for a discussion and examples). It is important to understand the ramifications of identity to both to limit the negative consequences and to be able to use one's sense of identity as a positive force in our world. We constructed a weak mechanism of group favoritism (color assignment, a form of the minimal-group paradigm) and augmented it with communication. Since pairs were switched half of the time, this initial communication is either direct (with no switch) or indirect (with a switch).

We test for differences in in-group versus out-group allocations when there is no direct communication and compare these differences to those found with direct communication. We confirm favoritism towards one's color group when there are no (mutual) promises to act favorably: Dictators are 11 percentage points more likely to sacrifice own payoffs when they are paired with same-color recipients than with different-color recipients, so our rather modest identity inducement was effective. In principle, there could be a null effect of communication on this gap, or even a negative one if communication tended to exacerbate differences or strength in-group ties. Mutual promises are highly effective, with favorable dictator choices made nearly 70% of the time.

What is surprising and gratifying is that the significant difference across recipient color-matches completely disappears when we consider only those cases where there were promises to *Roll*. Out-group and in-group promisors appear to be motivated by different mechanisms. As a result, the elimination of the ingroup bias seems to be driven primarily by the moral commitment that motivates outgroup promisors.

We also find the intriguing result that dictators are *more* likely to sacrifice when neither person in a pair makes a favorable declaration than when one person does, with roughly a 9-13 percentage-point difference in rates. Pairing across types of people may generate worse social outcomes than pairing two openly-selfish individuals if the more pro-social party is unable to persuade the other to promise to sacrifice. A caveat is that our statistical power is weakened by the need for so many sub-categories for this analysis, so that conservative tests do not confirm the significance of this difference. Nevertheless, this is a consideration for future study designs.

So, we not only find that promises lead to more willingness to make pro-social choices in both pairs that remained intact and those that did not, but we also find that these promises appear to transform strangers into fellows, since everyone is treated the same regardless of their color. We consider this to be a hopeful sign. Having direct communication here bridged the gap between strangers and partners. Perhaps this finding would also apply in field environments. We invite others to further investigate the interaction between communication and social identity.

References

Adnan, W., K. P. Arin, G. Charness, J.A. Lacomba, and F. Lagos (2022), "Which social categories matter to people: An experiment," *Journal of Economic Behavior and Organization*, 189: 125-145.

Akerlof, G. A. and R. E. Kranton (2000), "Economics and identity," *The Quarterly Journal of Economics* 115: 715–753.

Brandts, J., Charness, G. and M. Ellman (2016), "Let's talk: How communication affects contract design", *Journal of the European Economic Association*, 14: 943–974.

Charness, G. (2000a), "Self-serving cheap talk: A test of Aumann's conjecture", *Games and Economic Behavior*, 33: 177–194.

Charness, G. (2000b), "Bargaining Efficiency and Screening: An Experimental Investigation," *Journal of Economic Behavior and Organization*, 42: 285-304.

Charness, G., Ramón Cobo-Reyes and N. Jiménez (2014), Identities, Selection, and Contributions in a Public-goods Game," *Games and Economic Behavior*, 87: 322-338

Charness, G. and Y. Chen (2020), "Social identity, group behavior, and teams", *Annual Review of Economics*, 12: 691-713.

Charness, G., R. Cobo-Reyes and N. Jiménez (2014), "Identities, selection, and contributions in a public-goods game," *Games and Economic Behavior* 87: 322–338.

Charness, G. and M. Dufwenberg (2006), "Promises and partnership," *Econometrica,* 74: 1579–1601.

Charness, G. and M. Dufwenberg (2010), "Bare promises: An experiment", *Economics Letters,* 10: 281-283.

Charness, G., F. Feri, M. A. Meléndez-Jiménez, and M. Sutter (2021), "An experimental study on the effects of communication, credibility, and clustering in network games," forthcoming in *Review of Economics and Statistics*.

Charness, G. and P. Holder, (2018), "Charity in the laboratory: Matching, competition, and group identity," *Management Science,* 65, 1398-1407.

Charness, G., L. Rigotti and A. Rustichini (2007), "Individual behavior and group membership," *American Economic Review,* 97: 1340–1352.

Charness, G. and M. Villeval (2009), "Cooperation and competition in intergenerational experiments in the field and the laboratory", *American Economic Review,* 99: 956-78.

Chen, R. and Y. Chen (2011), "The potential of social identity for equilibrium selection," *American Economic Review,* 101: 2562–2589.

Chen, Y. and S. X. Li (2009), "Group identity and social preferences," *American Economic Review*, 99: 431–457.

Chen, Y., S. X. Li, T. Liu, and M Shih (2014), "Which hat to wear? Impact of natural identities on coordination and cooperation," *Games and Economic Behavior*, 84: 58-86.

Ciccarone. G., G. Di Bartolomeo and S. Papa (2020), "The rationale of in-group favoritism: An experimental test of three explanations", *Games and Economic Behavior,* 124: 554-568.

Cohn, A., T- Gesche, and M.A. Maréchal (2021), "Honesty in the digital age," *Management Science,* 68(2): 827-845.

Coffman, K.B. (2014), "Evidence on self-stereotyping and the contribution of ideas," *Quarterly Journal of Economics,* 129(4): 1625-1660.

Cooper, R., D.V. DeJong, R. Forsythe, and T.W. Ross (1989), "Communication in the Battle of the Sexes Game: Some Experimental Results", *The RAND Journal of Economics,* 20: 568-587.

Dahl, D.W., C. Fuchs, M. and Schreier (2014), "Why and when consumers prefer products of user-driven firms: A social identification account," *Management Science*, 61(8): 1978-1988.

Del Carpio, L. and M. Guadalupe (2021), "More women in tech? Evidence from a field experiment addressing social identity," *Management Science*, forthcoming.

Di Bartolomeo, G., M. Dufwenberg, S. Papa and F. Passarelli (2019a), "Promise, expectations and causation," *Games and Economic Behavior*, 113: 137–146.

Di Bartolomeo, G., M. Dufwenberg, and S. Papa (2019b), "The sound of silence: A license to be selfish," *Economics Letters*, 182: 68-70.

Di Bartolomeo, G., M. Dufwenberg, and S. Papa (2021), "Promises and Partner-Switch", *working paper*.

Fischbacher, U. (2007), "z–Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, 10: 171–178.

Güth, W., M. Ploner and T. Regner, (2009), "Determinants of in-group bias: Is group affiliation mediated by guilt-aversion?" *Journal of Economic Psychology*, 30: 814–827.

Hargreaves Heap, S. P. and D. J. Zizzo (2009). "The value of groups," *American Economic Review*, 99: 295–323.

Khalmetski, K., A. Ockenfels, and P. Werner (2015), "Surprising gifts: Theory and laboratory evidence," *Journal of Economic Theory*, 159: 163–208.

Kranton, R. and S. Sanders (2017), "Groupy vs. non groupy social preferences: Personality, region, and political party," *American Economic Review Papers and Proceedings*, 107: 65–69.

Kranton, R., M. Pease S. Sanders, and S. Huettel (2018), "Groupy vs. non-groupy behavior: Deconstructing group bias," Duke University, mimeo.

Niessen-Ruenzi, A. and Ruenzi S. (2019), "Sex matters: Gender bias in the mutual fund industry," *Management Science*, 65(7): 3001-3025.

Ockenfels, A. and P. Werner (2014), "Beliefs and in-group favoritism," *Journal of Economic Behavior & Organization*, 108: 453–462.

Reagans, R. (2005), "Preferences, identity, and competition: Predicting tie strength from demographic data," *Management Science*, 51(9): 1374-1383.

Shayo, M. (2009), "A model of social identity with an application to political economy: Nation, class, and redistribution," *American Political Science Review*, 103, No. 2.

Shih, M., T. L. Pittinsky, and N. Ambady (1999), "Stereotype Susceptibility: Identity Salience and Shifts in Quantitative Performance," *Psychological Science*, 10: 80-83.

Tajfel, H., and M. Billig (1973), "Social categorization and similarity in intergroup behaviour," *European Journal of Social Psychology*, 3: 27-52.

Tajfel, H. and J. C. Turner (1979), "An integrative theory of intergroup conflict," in W. G. Austin and S. Worchel (eds.), *The Social Psychology of Intergroup Relations*. Brooks/Cole, Monterey: 33–47.

Vanberg, C. (2008), "Why do people keep their promises? An experimental test of two explanations," *Econometrica*, 76: 1467–1480.

Appendix

Table A reports the second-order beliefs of dictators who did a promise by their switching condition within the ingroup and outgroup subsamples. As said, to get a fair comparison, we only take account of dictators re-matched with recipients who were previously outgroup (ingroup) when the outgroup (ingroup) comparison is considered.

As expected, within outgroup subjects, the second-order beliefs of non-switched dictators are not statistically different from those switched. Specifically, for outgroup dictators (row (1) vs. (2)), we find: 0.81 vs. 0.83: $Z = 0.81$, $p = 0.859$. Similarly, for ingroup dictators (row (3) vs. (4)), we find: 0.80 vs. 0.80: $Z = 0.31$, $p = 0.754$.

Table A. Average second-order beliefs of selected dictators who made a promise (718 obs.)

	Session												
Category	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**All**
Outgroup dictators													
(1) Non-Switched[(a)]	.86	.71	.77	.69	.77	.87	.91	.89	.80	.96	.92	.70	**.81**
(2) Switched[(b)]	.90	.58	1.0	.50	.81	.77	.92	.71	na	1.0	.77	.92	**.83**
Ingroup dictators													
(3) Non-switched[(c)]	.72	.83	.82	73	.81	.74	.71	.88	.92	.83	.85	.78	**.80**
(4) Switched[(d)]	.91	.79	.90	.79	.66	.80	.54	.93	.94	.75	.67	.75	**.80**

Notes: (a) Non-switched outgroup dictators who participated in a promise; (b) Switched outgroup dictators who participated in a promise and were rematched with a recipient who was also participating in promise with another outgroup partner. (d) Non-switched ingroup dictators who participated in a promise; (d) Switched ingroup dictators who participated in a promise and were rematched with a recipient who was also participating in promise with another ingroup partner.