Preferential interactions promote blind cooperation and informed defection

Alfonso Pérez-Escudero[a,1,2], Jonathan Friedman[a,1], and Jeff Gore[a]

[a]Physics of Living Systems, Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139

Edited by Matthew O. Jackson, Stanford University, Stanford, CA, and approved October 21, 2016 (received for review May 2, 2016)

It is common sense that costs and benefits should be carefully weighed before deciding on a course of action. However, we often disapprove of people who do so, even when their actual decision benefits us. For example, we prefer people who directly agree to do us a favor over those who agree only after securing enough information to ensure that the favor will not be too costly. Why should we care about how people make their decisions, rather than just focus on the decisions themselves? Current models show that punishment of information gathering can be beneficial because it forces blind decisions, which under some circumstances enhances cooperation. Here we show that aversion to information gathering can be beneficial even in the absence of punishment, due to a different mechanism: preferential interactions with reliable partners. In a diverse population where different people have different—and unknown—preferences, those who seek additional information before agreeing to cooperate reveal that their preferences are close to the point where they would choose not to cooperate. Blind cooperators are therefore more likely to keep cooperating even if conditions change, and aversion to information gathering helps to interact preferentially with them. Conversely, blind defectors are more likely to keep defecting in the future, leading to a preference for informed defectors over blind ones. Both mechanisms—punishment to force blind decisions and preferential interactions—give qualitatively different predictions, which may enable experimental tests to disentangle them in real-world situations.

cooperation | signaling | population heterogeneity | game theory | incomplete information

Why didn't you ask before? This question too often lacks a reasonable answer. Consider the case of Alice who, learning that her friend Bob was visiting her city, rushed to invite him to stay at her place—only to find, once the offer had been gratefully accepted, that Bob would occupy her living room for a whole month. Most people would agree: Alice should have asked before.

However, Alice's behavior is not uncommon. We often refrain from gathering information about the costs and benefits of our interactions with others, and often with good reason: When the situation is reversed, we prefer people who directly agree to do us a favor over those who only agree after carefully weighing costs and benefits. Why do we have this preference? If someone agrees to do us a favor—or to cooperate with us in any other way—why should we care about how they made their decision?

Intuitively, aversion to information gathering can be beneficial because it makes us prefer people who are more likely to cooperate. In addition, people who would in principle want to gather information may be inhibited by others' aversion to it. To illustrate these two mechanisms, let us first consider a simple scenario in which two types of people exist: unreliable and reliable. Unreliable types only cooperate in a limited set of conditions and need to gather additional information to decide whether they will cooperate in any particular situation. In contrast, reliable types cooperate in all situations. Therefore, those who gather information reveal themselves as unreliable, and aversion to information gathering allows us to avoid them (Fig. 1*A*). In contrast, consider now a situation in which only the unreliable type exists. Here information gathering does not inform about the type, because there is only one (the same would happen in a population with several types, if we can distinguish them in any other way). However, the unreliable cooperators can sometimes be manipulated: The threat of a punishment for gathering information may force them to make a blind decision, which under some conditions will be more cooperative than an informed one. Aversion to information gathering can therefore help to manipulate these unreliable types (Fig. 1*B*).

Both preferential interactions and manipulation are well understood, but not in the context of information gathering. Manipulation (in a broad sense) is at the core of game theory, where one player's strategy may determine the other's (1, 2). In particular, cooperation is often sustained by the threat of a future punishment or the promise of a future reward (3–6). Preferential interactions are beneficial in heterogeneous populations, where long-lasting relationships are formed with beneficial partners and not with detrimental ones (7–10). Although distinct, both mechanisms are often interdependent, for example when punishment takes the form of terminating a beneficial relationship (7–10). Preferential interactions are often studied in the context of incomplete information, where signals indicate (explicitly or implicitly) whether their sender is a beneficial partner (7–12). However, none these studies addresses a player's choice to gather external information. Instead, they typically consider that one player has information that is relevant for the other (or for both) and may choose how much to share. The choice to gather information has been studied mainly in two cases: when gathering or processing information is costly, so subjects choose their actions using simple heuristics instead of fully analyzing the situation (13–15), and in the context of strategic ignorance, where

Significance

Humans often behave in seemingly irrational ways. A common instance of such perplexing behavior is that we typically care about how and why people chose their actions, rather than caring only about the actions themselves. For example, when people agree to do us a favor, we prefer them to do so directly, rather than to first gather all the relevant information. Using game theory, we show that this preference may in fact be rational: The decision-making process often reveals hidden preferences of the decision maker, which can become relevant in a future interaction. This work elucidates the conditions that make caring about motivations beneficial and makes predictions regarding the real-world situations in which it is expected to occur.

Author contributions: A.P.-E., J.F., and J.G. designed research; A.P.-E. and J.F. performed research; and A.P.-E., J.F., and J.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]A.P.-E. and J.F. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: alfonso.perez.escudero@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1606456113/-/DCSupplemental.

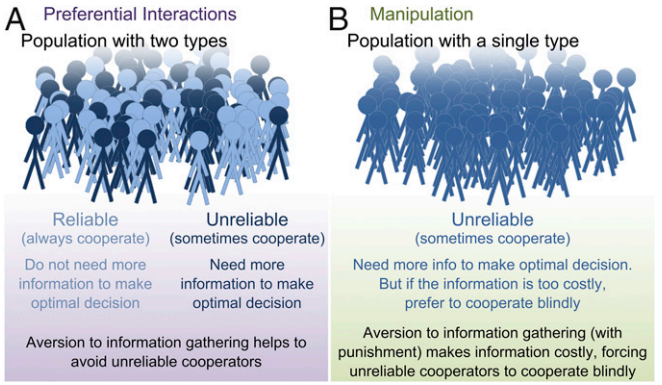

Fig. 1. Two mechanisms promote aversion to information gathering. (*A*) Preferential interactions: In a heterogeneous population with two types (reliable and unreliable), those who gather information reveal themselves to be unreliable, even if they cooperate in the present situation. Aversion to information gathering then helps to avoid these unreliable types. (*B*) Manipulation: Even if there is no uncertainty about the type (in this example because only one type exists), aversion to information gathering can be useful, because punishing information gatherers may prevent them from acquiring information. Without information, they may be forced to cooperate more often than they would otherwise.

having a given piece of information can—surprisingly—lead to lower payoffs (16–19). In contrast, we focus on the case where one player can costlessly acquire a piece of information that only refers to her own payoffs, and that would help her to make a more profitable decision.

This case has been addressed recently by a model called "the envelope game" (20, 21). In this paper we first show that—although both manipulation and preferential interactions are usually invoked to explain aversion to information gathering (20–24)—the current implementation of the envelope game only captures manipulation. Next, we extend the framework to heterogeneous populations to include preferential interactions. We find qualitative differences between the two mechanisms, with experimentally testable differences about the effect of punishment and the cost of detecting information gathering. Finally, we show that in some conditions preferential interactions lead to the opposite effect: preference for information gathering.

Results

The Envelope Game in Homogeneous Populations. Let us begin with the original implementation of the envelope game (20), which is a repeated asymmetric game between two players (Fig. 2*A*). In every round, player 1 receives a closed envelope that contains a temptation to defect. This temptation can have a low value ($c_l > 0$) with probability p, or a high value ($c_h > c_l$) with probability $1-p$. Player 1 must first decide whether to look inside the envelope and then whether to cooperate or to defect. If player 1 cooperates, she gets a payoff a and player 2 gets a payoff b; if player 1 defects, she gets whatever was in the envelope (either c_l or c_h) and player 2 gets a negative payoff, $d < 0$. Then, player 2, who is aware of player 1's actions and of the content of the envelope, decides whether to continue or to end the game. If he decides to continue, another round is played with probability w (with probability $1-w$, the game ends anyway). In the new round, a new temptation is randomly put inside the envelope. We will consider the following strategies, which will be sufficient to describe our main results: Player 1 can (*i*) cooperate without looking, (*ii*) cooperate with looking (meaning that she will look in the envelope but then will always cooperate regardless of the temptation), (*iii*) look in the envelope and cooperate if the temptation is low and defect if it is high, or (*iv*) always defect. Player 2 can (*i*) always end the game, regardless of player 1's behavior, (*ii*) end the game if player 1 defects, or (*iii*) end the game if player 1 looks or defects (we abbreviate this strategy with "end if P1 looks"). Each player's average payoff depends both on the payoffs distributed in each round and on the number of rounds played (Fig. 2*B*). Except for the last section, where we will discuss preference for looking, we will assume that $pb + (1-p)d < 0$, meaning that player 2 is harmed on average by a player 1 who only cooperates when the temptation is low.

As in the real-life situations the game intends to mimic, in any given round the payoffs depend only on whether player 1

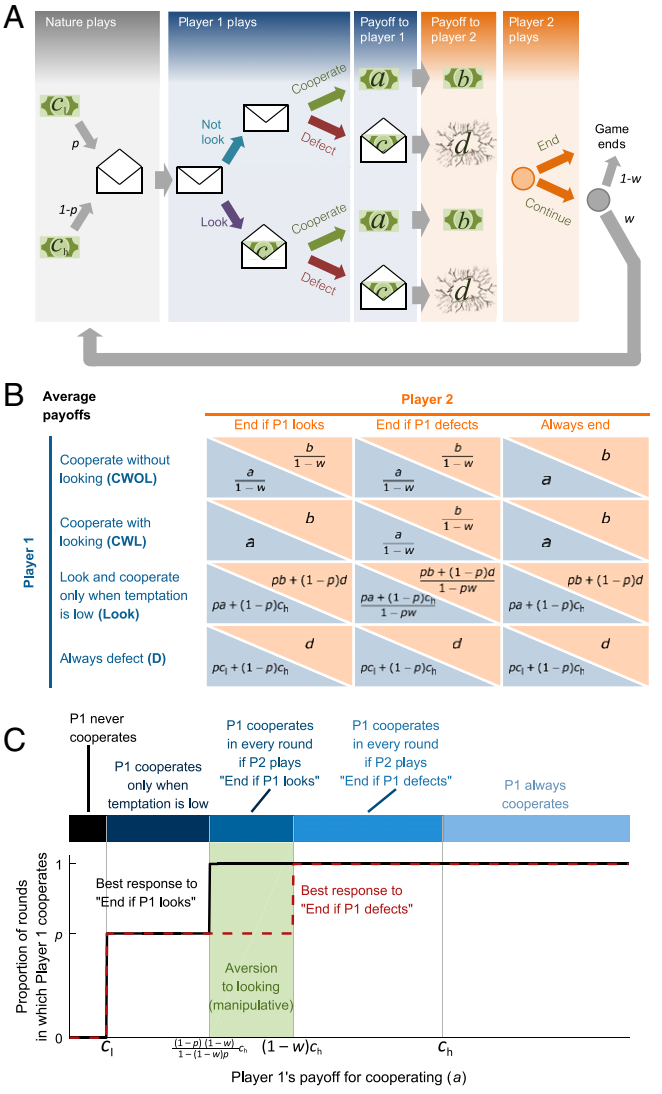

Fig. 2. In homogeneous populations, the envelope game predicts aversion to looking as a means to manipulate the other player's strategy. (*A*) Schematic of the game. (*B*) Payoff table for the essential strategies. Strategy "end if P1 looks" stands for "end if P1 looks or defects." All payoffs are long-term averages [for example, the top-left payoff to player 1 is the single-round payoff a times the average number of rounds $(1-w)^{-1}$]. Payoffs for player 1 on blue background, and payoffs for player 2 on orange background. (*C*) Proportion of rounds in which player 1 cooperates, when playing a best response against "end if P1 looks" (solid black line) or "end if P1 defects" (red dashed line), and as a function of parameter a (payoff to player 1 when she cooperates). The rest of the parameters are fixed at $b = 1.5$, $c_l = 1$, $c_h = 10$, $d = -5$, $w = 0.4$, and $P = 0.6$. One can check that the conditions in the x axis are necessary for each best response by inspecting the payoff table in *B*. *SI Appendix*, section 2 shows that they are also sufficient and gives more general expressions. Green shade marks the region where aversion to looking exists.

cooperates or defects, regardless of whether she looks in the envelope or not. Therefore, why should player 2 care about looking, and not just about cooperation? To identify conditions in which caring about looking is beneficial to player 2 we explore the Nash equilibria of the game, which are states in which no player has an incentive to deviate from their current strategy. We say that aversion to looking exists when "cooperate without looking" and "end if P1 looks" form a Nash equilibrium, whereas "cooperate with looking" and "end if P1 defects" do not. We exclude conditions where this second equilibrium exists because all payoffs are identical in both equilibria, so in these situations caring about looking is not strictly beneficial to player 2.

Hoffman et al. (20) showed that aversion to looking exists when player 2 can enforce reliable cooperation by preventing player 1 from looking. To illustrate this, let us consider a case in which player 1's payoff for cooperation is between the two temptations to defect, $c_l < a < c_h$. To maximize her one-round payoff, she should play "look and cooperate only when temptation is low." However, if player 2 plays "end if P1 defects," defecting to collect the high temptation means losing the benefits of all future rounds. She will therefore prefer to cooperate in every round if these future benefits are higher than the high temptation [$a/(1-w) > c_h$, where $1/(1-w)$ is the average number of rounds; Fig. 2*C*, red dashed line]. In contrast, if player 2 plays "end if P1 looks," player 1 must decide without knowing the exact temptation. Her expected payoff if she looks is $pa + (1-p)c_h$, so she will cooperate without looking when $a/(1-w) > pa + (1-p)c_h$ (Fig. 2*C*, black line). Aversion to looking exists in the intermediate range where "end if P1 looks" forces player 1 to cooperate whereas "end if P1 defects" does not (Fig. 2*C*; see *SI Appendix*, section 2 for the general proof). Therefore, aversion to looking is always manipulative in homogeneous populations.

It is illustrative to consider this result in the context of our introductory host's dilemma: Alice (player 1) would benefit from asking Bob (player 2) for how long he wants to stay, and host him only if the visit is short. However, if Bob reacts to the question by punishing Alice, for example by getting angry or terminating the friendship (the equivalent of ending the game), Alice will refrain from asking. Bob benefits from this situation: He can abuse Alice's hospitality with occasional long stays, because she will always agree to host him without asking. This mechanism requires Bob to be able to punish Alice strongly enough, a requirement that is often not met in real life. We will now show that this mechanism—which emerges from the current implementation of the envelope game (20)—is only one of the possible explanations for the existence of aversion to information gathering.

The Envelope Game in Heterogeneous Populations. In real life, different people face different costs and benefits when facing the same situation. For example, the cost of hosting a friend depends on whether one lives in a big house with a guest room or in a small studio, and the pleasure derived from the visit depends on how much one likes the visitor. Even when interacting with people we know, some of these factors remain uncertain, which makes it difficult to predict the other person's preferred strategy. This uncertainty is not reflected in the envelope game as described above, which is a game with imperfect but complete information (imperfect because the temptation in the envelope is unknown; complete because both players have the same information).

A simple way to incorporate the uncertainty about the other player's preferences is to assume that the population is heterogeneous, containing different types of players with different payoffs (25, 26). In the simplest case we consider a population composed of two types of player 1, which differ only in their payoff for cooperation (parameter a). We call the type with higher a "favorable type" (or player 1_F), because she will be more prone to cooperating. Similarly, we call the type with lower a "unfavorable type" (or player 1_U). Player 2 knows that there are two types and the differences between them but does not know with which type he is playing in any given interaction, so the game now has incomplete information. Player 2 must find a strategy that maximizes his expected payoff in every situation, given his best estimate about the type he is playing with [in technical terms, we are interested in the sequential equilibria of the game (27, 28)].

In such a heterogeneous population, aversion to looking occurs in a wide range of conditions and does not necessarily imply manipulation (Fig. 3*A*): Aversion to looking can exist even when "end if P1 looks" fails to change the behavior of any player 1 (Fig. 3*B*, *Top*). In these conditions, aversion to looking arises because it helps player 2 to interact preferentially with the favorable type, who cooperates in every round and does not need to look. The unfavorable type looks and cooperates only when the temptation is low, so "end if P1 looks" allows prolonged beneficial interactions with the favorable type, while ending the game immediately when playing with the unfavorable one. In contrast, "end if P1 defects" is not equally efficient at protecting player 2 from the unfavorable type (Fig. 3*B*, *Bottom*).

Thus, population heterogeneity introduces a new mechanism that supports aversion to looking: preferential interactions with the favorable types in the population. The manipulative mechanism that we saw in homogeneous populations is also present in heterogeneous ones, taking place whenever at least one type of player 1 can be forced to cooperate by "end if P1 looks" but not by "end if P1 defects" (green and striped areas in Fig. 3 *A*, *C*, and *D*). Each mechanism on its own can produce aversion to looking, but they can also act simultaneously (striped areas in Fig. 3 *A* and *C*). See *SI Appendix*, section 2 for the general conditions where each mechanism takes place.

We find no evidence of any other mechanism: Whenever aversion to looking exists, it is due either to manipulation, preferential interactions, or both (proof is given in *SI Appendix*, section 2).

The mechanism of preferential interactions adds a new interpretation to the host's dilemma: Bob is unsure about how willing Alice is to host him (represented by a in the model). Alice's question—or its absence—gives Bob extra information: If she needs to ask to make the decision, she is probably not very happy to host him (in the model, looking indicates that a is below some threshold). Friends who ask thus reveal themselves to be less prone to cooperate than those who do not, and although this may not matter in the current interaction, it may matter in the future. Thus, Bob should prefer to interact with friends who do not ask.

Our analysis has been based on the existence of a sequential equilibrium where player 1 plays "cooperate without looking" and player 2 plays "end if P1 defects." Even when this equilibrium exists, the game may have other alternative equilibria (see *SI Appendix*, section 7). To test that the equilibrium that gives rise to aversion to looking is dynamically stable, we have run simulations using the replicator equation. We find that the equilibrium is indeed stable and is chosen over alternatives with high probability, especially in heterogeneous populations (*SI Appendix*, section 3).

In the following sections we show that preferential interactions increase the prevalence of aversion to looking, and that they differ from the manipulative mechanism in several qualitative ways: (*i*) They remove the need for punishment, (*ii*) they make aversion to looking robust to costly detection of looking, and (*iii*) they may create the opposite effect: preference for looking.

Prevalence of Aversion to Looking in Heterogeneous Populations. Population heterogeneity promotes aversion to looking by widening the conditions in which it exists. Increasing population heterogeneity by adding a new type of player 1 can only enable an aversion to looking, and never prevent it (see *SI Appendix*, section 2). The intuition behind this result is that the strategy

"end if P1 looks" works well with any type of player 1: It protects player 2 from types with low values of a by ending the game at once, forces types with intermediate values of a to cooperate, and maintains the interaction with types with high values of a. Therefore, adding new types of player 1 increases the probability of hitting a combination of types that requires playing "end if P1 looks." Likewise, in populations with many types the strength of aversion to looking increases as the types become more different (Fig. 4; see *SI Appendix*, section 2 for a general proof).

The Role of Punishment. In the envelope game, when player 2 ends the game he simultaneously protects himself from potential future losses and strips player 1 from potential future gains, effectively punishing her. This punishment is required for the manipulative mechanism that leads to aversion to looking in homogeneous populations. In contrast, the mechanism of preferential interactions in heterogeneous populations does not need any kind of punishment: Aversion to information gathering can emerge even if player 2 cannot affect player 1's payoffs—that is, even if player 1's payoffs are the same regardless of whether player 2 ends the game (*SI Appendix*, section 4).

Aversion to Looking When Detecting Looking Is Costly. In real-life situations it can be costly to detect that another person is gathering information. If player 2 needs to pay such a cost—even an infinitesimal one—the strategy "end if P1 looks" cannot be part of a pure equilibrium in a homogeneous population (*SI Appendix*, section 5). Although it can be part of a mixed equilibrium, mixed equilibria are not stable in asymmetric games such as the envelope game (29, 30). Therefore, this cost would make aversion to looking unsustainable in homogeneous populations. In contrast, the cost makes no qualitative difference in a heterogeneous population: Aversion to looking remains stable for as long as the benefit it gives to player 2 outweighs the cost (*SI Appendix*, section 5).

Preference for Looking. So far we have discussed aversion to looking. However, in some conditions the model predicts the opposite effect: Player 2 will end the game if player 1 does not look. The transition between aversion to looking and preference for looking depends on whether player 2 obtains a net negative or positive payoff from interacting with unreliable cooperators (types of player 1 who look in the envelope and cooperate only when the temptation to defect is low). Until now we have only studied the case when this net payoff was negative $[pb + (1-p)d < 0]$, and aversion to looking protected player 2 from the unreliable cooperators.

When the interactions with unreliable cooperators are beneficial $[pb + (1-p)d > 0]$, player 2 no longer needs to protect himself from them. In this situation, the problem is to distinguish unreliable cooperators (who are beneficial) from all-defectors (who are detrimental). Looking is informative because a player 1 who defects without looking is surely an all-defector, whereas one who looks before defecting may be an unreliable cooperator.

To analyze preference for looking we need to consider new strategies: for player 1, "defect with looking" and "defect without looking"; for player 2, "end if P1 defects when temptation is low or defects without looking" and "end if P1 defects when temptation is low (regardless of looking)" (*SI Appendix*, section 6). We say that preference for looking exists when there is a sequential equilibrium where at least one type of player 1 plays "defect with looking" and player 2 plays "end if P1 defects when temptation is low or defects without looking," and there is no payoff-equivalent sequential equilibrium where a type of player 1

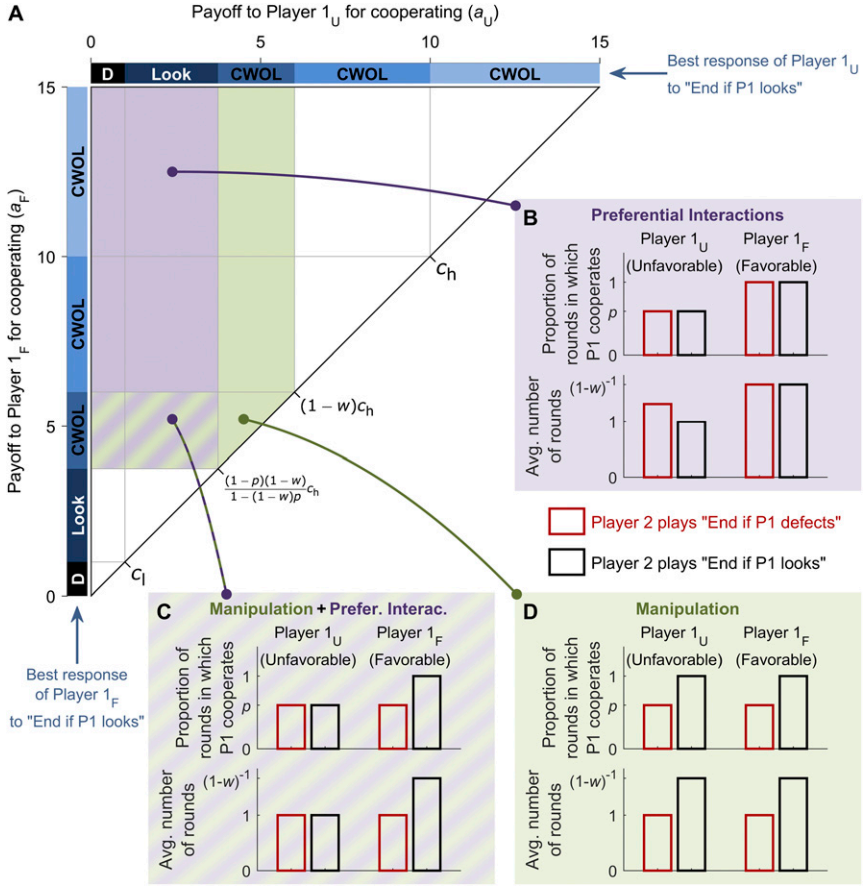

Fig. 3. In heterogeneous populations, aversion to looking allows player 2 to interact preferentially with beneficial partners. (*A*) Existence of aversion to looking in a population with two types, as a function of their payoffs for cooperation (a_U for the unfavorable type and a_F for the favorable one). The rest of the parameters are fixed at $b = 1.5$, $c_l = 1$, $c_h = 10$, $d = -5$, $w = 0.4$, and $P = 0.6$. White: aversion to looking does not exist. Green: aversion to looking exists and is purely manipulative. Purple: aversion to looking exists, and comes purely from preferential interactions. Stripes: aversion to looking exists, and both mechanisms are at play simultaneously. Blue bars indicate the best response of each type of player 1 to "end if P1 looks" for each value of parameter a (CWOL, "cooperate without looking"; D, "always defect"; Look, "look and cooperate only when temptation is low"). (*B*) Analysis for a representative point where aversion to looking comes purely from preferential interactions. (*Top*) Proportion of rounds in which each type of player 1 cooperates when playing a best response to "end if P1 defects" (red) or to "end if P1 looks" (black). (*Bottom*) Average number of rounds played against each type of player 1 when player 2 plays "end if P1 defects" (red) or "end if P1 looks" (black), and player 1 plays her best response. (*C*) Same as *B*, but for a representative point of the region with both manipulation and preferential interactions. (*D*) Same as *B*, but for a representative point of the region where only preferential interactions take place.

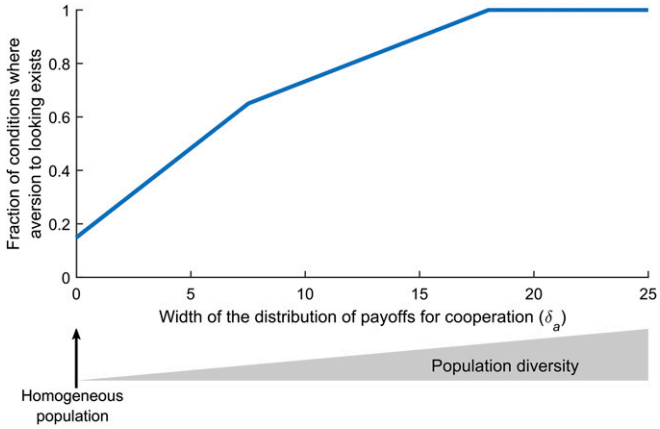

Fig. 4. Population diversity promotes aversion to looking. We consider a population with infinite types, whose payoff for cooperation (a) is distributed uniformly in an interval of width δ_a. We show the proportion of cases in which aversion to looking exists as a function of the width of the interval, averaged over all possible centers of the interval between 0 and 15. The rest of the parameters are fixed at $b = 1.5$, $c_l = 1$, $c_h = 10$, $d = -5$, $w = 0.4$, and $P = 0.6$. For the analytical derivation of the line, see *SI Appendix*, section 2.

plays "defect with looking" and player 2 plays "end if P1 defects when temptation is low (regardless of looking)."

Preference for looking requires population heterogeneity and emerges when at least one type of player 1 always defects, at least one type of player 1 looks and cooperates when the temptation is low, and when the interactions with the unreliable type are very beneficial for player 2. To illustrate this point, we consider again a population with two types of player 1 (X and Y). We fix the payoff for cooperation of type Y (a_Y), such that she will be an unreliable cooperator (i.e., will play "look and cooperate if the temptation is low"). Then, we study the game when changing the payoff for cooperation for the other type (a_X), and also the payoff that player 2 obtains when player 1 cooperates (b). This second parameter controls whether the interactions with unreliable cooperators are on average detrimental or beneficial for player 2. When these interactions are detrimental, we recover our previous results: Aversion to looking emerges when type X has high enough payoff to cooperate without looking. Preference for looking emerges when type Y has a very low payoff for cooperation, so will always defect, and when the interactions with the unreliable type are very beneficial for player 2 (Fig. 5*C*). In the intermediate regime, neither of them arises: Interactions with unreliable types are beneficial, so aversion to looking makes no sense, but they are not beneficial enough to compensate for the possibility of being tricked by all-defectors who look. In this regime other strategies such as "end if P1 defects" are most efficient, and player 2 is indifferent to looking. See *SI Appendix*, section 6 for the general description.

Real-life situations may contain both aversion and preference for information gathering. For example, when somebody asks for an unspecified favor, the kindest answer is to agree without hesitation (i.e., not gather information). However, in many cases it is acceptable to ask what favor it is before granting it. What is not acceptable is to refuse to do the favor without even asking what it was. This response is usually considered very rude, and in fact it is used (very infrequently) to actively signal dislike toward the person who asked for the favor. The envelope game contains the essentials to understand this situation. First, refusing without asking about the favor signals that even low-cost favors would not be granted. Second, we rarely see this response, because mimicking a willingness to cooperate by asking for details is practically costless. Third, because of this low cost we consider asking before refusing as the baseline, and not doing so as an active signal.

Discussion

We conclude that two different mechanisms generate aversion to information gathering: manipulation, where the aversion enforces blind cooperation, and preferential interactions, where the aversion distinguishes between desirable and undesirable partners. In homogeneous populations we can only find the manipulative mechanism, whereas heterogeneous populations support both mechanisms.

Our results widen the conditions where we can expect aversion to looking, adding new qualitative predictions. From a model incorporating only manipulation, we would expect aversion to looking to occur only when punishment is strong enough to enforce cooperation, whereas our results show that punishment is often not required at all. Also, pure manipulation would not predict aversion to looking if detection of looking is costly, whereas preferential interactions do. Finally, preferential interactions give rise to the opposite effect: preference for information gathering. These differences between the two mechanisms are experimentally testable, for example by manipulating the strength of punishment and the cost of detecting information gathering in controlled experiments (such as those in refs. 23 and 24). Also, when analyzing real-life situations manipulation and preferential interactions will occur simultaneously, and they will be difficult to disentangle. The role of punishment and cost of detection of information gathering may give us a way to distinguish the dominant mechanism in different situations. For example, lack of punishment indicates preferential interactions, whereas strong punishment indicates manipulation. In particular, spiteful behavior, where individuals are willing to pay a cost to punish others, is suggestive of manipulation.

A general principle encompasses both aversion and preference for information gathering. Information gathering indicates that the current behavior (either cooperation or defection) may change in the future. Therefore, when people cooperate we prefer them to do so without gathering information; when they defect, we prefer them to do so after gathering information.

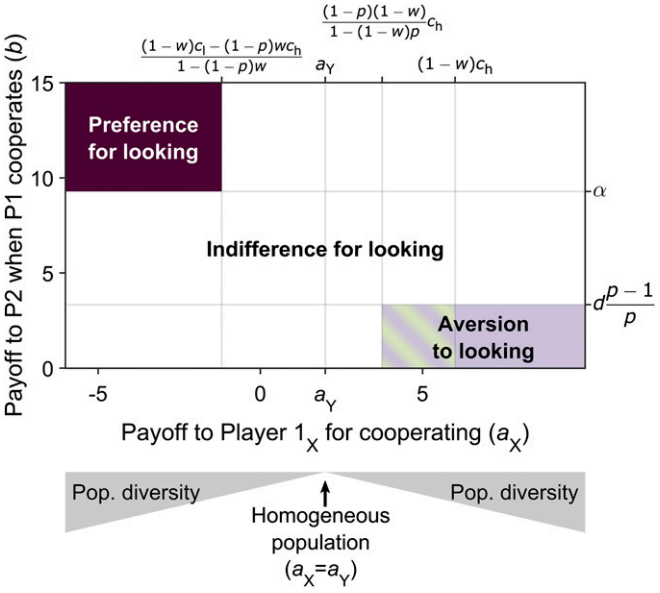

Fig. 5. Transition between aversion to looking and preference for looking. Occurrence of aversion to looking (purple and green/purple stripes) and preference for looking (dark red) in a population with two types of player 1 (X and Y). Type Y has fixed payoff for cooperation, $a_Y = 3$, and the x axis shows the payoff for cooperation of type X (a_X). The y axis shows the payoff that player 2 gets when player 1 cooperates (b). The rest of the parameters are fixed at $b = 1.5$, $c_l = 1$, $c_h = 10$, $d = -5$, $w = 0.4$, and $P = 0.6$. $\alpha = -d(2-p-w-(1-p)^2w)/[p-(1-p)pw]$. See *SI Appendix*, section 6 for the derivations.

Preference for looking is closely related to excuses. Consider a modification of the game where player 2 cannot see the temptation in the envelope, and player 1 can lie about it. Then, an all-defector could adopt the strategy "look in the envelope, always report the high temptation and defect," to appear as an unreliable cooperator. This behavior is comparable to giving an excuse, where one signals a cost for cooperation higher than the actual one.

Manipulative mechanisms in repeated games are often complicated by renegotiation, if the players can communicate and reconsider their strategies during the game. For example, consider a player 1 who cooperates without looking in response to player 2's threat to end the game. If player 1 deviates and defects, player 2 should end the game. However, both players would benefit from player 2's forgiving this deviation, if player 1 reverts to cooperate without looking. This situation produces a conflict. On the one hand, a credible threat of punishment is needed to force player 1 to cooperate. On the other, forgiveness is beneficial for both players after the actual deviation. This conflict is present in the envelope game when only the manipulative mechanism is active [in this case, the equilibrium is not renegotiation-proof (31, 32)]. A detailed study of this issue may be an interesting topic for future research, especially because pure preferential interactions are less prone to renegotiation: When playing against an unreliable cooperator who cannot be manipulated, renegotiation plays no role because player 1 would never agree to cooperate without looking, and when playing against a reliable cooperator, punishment is not needed.

The envelope game may have more equilibria than those discussed here. In particular, in some conditions there is a sequential equilibrium where player 2 always end the game (*SI Appendix*, section 7). This equilibrium always coexists with aversion to looking in homogeneous populations, and in some conditions in heterogeneous ones. However, whenever aversion to looking exists, the equilibrium where player 2 plays "end if P1 looks" Pareto-dominates the one where he plays "always end" (i.e., all players get higher or equal payoffs in the "end if P1 looks" equilibrium and for at least one player they are strictly higher; *SI Appendix*, section 7). Therefore, if given the opportunity to choose among both equilibria, all players would agree to choose the "end if P1 looks" one, which gives rise to aversion to looking.

The mechanism of preferential interactions is related to ideas regarding different levels of communication considered in psychology (33). These argue that messages can convey both their explicit content and implicit information about the relationship of the two actors (something referred to as "report" and "command" aspects of communication, respectively). In the context of our model, the explicit message can be player 1's question asking for details (for example, how long the guest wants to stay). The implicit information comes from the fact that player 2 can refine his estimate about the value of a, which determines the relationship between both players (in the sense that it will determine the strategy of player 1 in future interactions).

In the envelope game, partner quality is inferred solely based on behavior during the current game. However, in reality, the partner's behavior during past interactions with others often provides further information regarding their quality, giving rise to reputation effects (34). In particular, seeking information can lead potential future partners to infer that the information gatherer is a less desirable partner, who is likely to cooperate only under a limited set of conditions, and therefore decline future interactions with him (23). Even when this avoidance is not intended to punish unreliable partners, it can still reduce their expected future payoffs and thus serve as an effective punishment that can enforce blind cooperation. Therefore, although the conditions for manipulation are restrictive, it may be prevalent in heterogeneous populations due to the impact of reputation on preferential interactions, highlighting the importance of taking population heterogeneity into account.

ACKNOWLEDGMENTS. We thank Muhamet Yildiz, Moshe Hoffman, Erez Yoeli, Christian Hilbe, Sara Arganda, Adela Pérez-Escudero, Barrett Deris, Nic Vega, Shreyas Gokhale, the members of the J.G. laboratory, and three anonymous reviewers for discussions and comments on the manuscript. This work was supported by EMBO Postdoctoral Fellowship Grant ALTF 818-2014, Human Frontier Science Foundation Postdoctoral Fellowship Grant LT000537/2015, and the Paul Allen Family Foundation.

1. Osborne MJ (2000) *An Introduction to Game Theory* (Oxford Univ Press, Oxford).
2. Maynard Smith J (1982) *Evolution and the Theory of Games* (Cambridge Univ Press, Cambridge, UK).
3. Fudenberg D, Maskin ES (1986) The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* 54(3):533–554.
4. Nowak MA (2006) Five rules for the evolution of cooperation. *Science* 314(5805): 1560–1563.
5. Axelrod R (1984) *The Evolution of Cooperation* (Basic Books, New York).
6. Trivers RL (1971) The evolution of reciprocal altruism. *Q Rev Biol* 46(1):35–57.
7. Ghosh P, Ray D (1996) Cooperation in community interaction without information flows. *Rev Econ Stud* 63:491–519.
8. Kranton RE (1996) The formation of cooperative relationships. *J Law Econ Organ* 12(1):214–233.
9. Rob R, Yang H (2010) Long-term relationships as safeguards. *Econ Theory* 43(2): 143–166.
10. Hauk E (2003) Multiple prisoner's dilemma games with(out) an outside option: An experimental study. *Theory Decis* 54:207–229.
11. Crawford VP, Sobel J (1982) Strategic information transmission. *Econometrica* 50(6): 1431–1451.
12. Sobel J (1985) A theory of credibility. *Rev Econ Stud* 52(4):557–573.
13. Kahneman D (2003) A perspective on judgment and choice: Mapping bounded rationality. *Am Psychol* 58(9):697–720.
14. Bear A, Rand DG (2016) Intuition, deliberation, and the evolution of cooperation. *Proc Natl Acad Sci USA* 113(4):936–941.
15. Rand DG, et al. (2014) Social heuristics shape intuitive cooperation. *Nat Commun* 5: 3677.
16. Schelling TC (1956) An essay on bargaining. *Am Econ Rev* 46(3):281–306.
17. Conrads J, Irlenbusch B (2013) Strategic ignorance in ultimatum bargaining. *J Econ Behav Organ* 92:104–115.
18. Poulsen AU, Tan JHW (2007) Information acquisition in the ultimatum game: An experimental study. *Exp Econ* 10(4):391–409.
19. McGoey L (2012) The logic of strategic ignorance. *Br J Sociol* 63(3):553–576.
20. Hoffman M, Yoeli E, Nowak MA (2015) Cooperate without looking: Why we care what people think and not just what they do. *Proc Natl Acad Sci USA* 112(6): 1727–1732.
21. Hilbe C, Hoffman M, Nowak M (2015) Cooperate without looking in a non-repeated game. *Games* 6(4):458–472.
22. Barclay P (2016) Biological markets and the effects of partner choice on cooperation and friendship. *Curr Opin Psychol* 7:33–38.
23. Jordan JJ, Hoffman M, Nowak M, Rand DG (2016) Uncalculating cooperation is used to signal trustworthiness. 113(31):85658–63.
24. Capraro V, Kuilder J (2016) To know or not to know? Looking at payoffs signals selfish behavior, but it does not actually mean so. *J Behav Exp Econ*, dx.doi.org/10.1016/j.socec.2016.08.005.
25. Rosenthal R (1979) Sequences of games with varying opponents. *Econometrica* 47(6): 1353–1366.
26. Harsanyi JC (2004) Games with incomplete information played by "Bayesian" players, I-III: Part I. The basic model. *Manage Sci* 50:1804–1817.
27. Kreps DM, Wilson R (1982) Sequential equilibria. *Econometrica* 50(4):863–894.
28. Osborne MJ, Rubinstein A (1995) *A Course in Game Theory* (MIT Press, Cambridge, MA).
29. Selten R (1980) A note on evolutionarily stable strategies in asymmetric animal conflicts. *J Theor Biol* 84(1):93–101.
30. Cressman R, Tao Y (2014) The replicator equation and other game dynamics. *Proc Natl Acad Sci USA* 11(Suppl 3):10810–10817.
31. Farrell J, Maskin E (1989) Renegotiation in repeated games. *Games Econ Behav* 1(4): 327–360.
32. Bernheim B, Ray D (1989) Collective dynamic consistency in repeated games. *Games Econ Behav* 1:295–326.
33. Watzlawick P, Beavin JH, Jackson DD (1967) *Pragmatics of Human Communication: A Study of Interactional Patterns, Pathologies, and Paradoxes* (Norton, New York), pp 51–54.
34. Nowak MA, Sigmund K (2005) Evolution of indirect reciprocity. *Nature* 437(7063): 1291–1298.