Evolution and Human Behavior 34 (2013) 258–264

Contents lists available at SciVerse ScienceDirect

Evolution and Human Behavior

journal homepage: www.ehbonline.org

Original Article

Can we see inside? Predicting strategic behavior given limited information

Sonja Vogt [a,b,*], Charles Efferson [a,b,*], Ernst Fehr [a,b,*]

[a] Department of Economics, University of Zurich, 8006 Zurich, Switzerland
[b] Laboratory for Social and Neural Systems Research, University of Zurich, 8006 Zurich, Switzerland

ARTICLE INFO

Article history:
Initial receipt 2 October 2012
Final revision received 18 March 2013

Keywords:
Cooperation
Coordination
Thin slices
Prisoner's dilemma
Stag hunt

ABSTRACT

Evolutionary theory predicts that observable traits should evolve to reliably indicate unobservable behavioral tendencies in coordination games but not social dilemmas. We conducted a two-part study to test this idea. First, we recorded 60-s videos of participants, and then these participants played a stag hunt game or a prisoner's dilemma. Subsequently, raters viewed these videos, with the sound either off or on, and they guessed player choices. Raters showed a significant tendency to guess that attractive players chose stag. In contrast to the prediction, rater accuracy was at chance regardless of whether the sound of the video was off or on. For prisoner's dilemma players, raters showed a significant tendency to guess that women cooperated at a higher rate than men. Again in contrast to the prediction, accuracy was significantly above chance in this case. To calibrate the importance of this accuracy rate, we developed two models that suggest the accuracy we observed in the prisoner's dilemma case is probably not high enough to support the evolution of cooperation. Altogether, our results show that raters tried to achieve a meaningful degree of accuracy about players by using the limited information available in the videos, but they could not do so.

1. Introduction

In strategic situations, what a person does will often depend on what she thinks about the people around her. Coordination games provide especially clear examples of this principle (Young, 1996). Coordination games have multiple equilibria, and players face at least partially congruent incentives to coordinate their behaviors. This is why beliefs about others matter. When Charles talks to his father, a native of South Louisiana, he occasionally uses the word "lagniappe." When he talks to his friend Ryan, a native of Waroona, Western Australia, he does not. Charles and his father both know that when they are together the word lagniappe will lead to coordinated communication. In contrast, Ryan is not a student of South Louisiana dialects, and Charles does not use them around him. Everyone has accurate expectations, and everyone is in equilibrium, though the equilibria depend on who is interacting with whom.

Beliefs about others can also prove crucial in social dilemmas. Social dilemmas and the associated evolution of cooperation stand as one of the most active and controversial areas of research in the study of human social behavior (Bowles & Gintis, 2011; Henrich, 2004). Social dilemmas have dominant strategies, which simply mean strategies that are optimal regardless of what others do. More specifically, in social dilemmas a decision maker can cooperate and, at some personal cost, produce benefits enjoyed by others. Because cooperation is individually costly, the dominant strategy is to defect unconditionally. This does not sound like a situation in which beliefs about others should matter. Defection is dominant, and seemingly this should be true whatever type of person one happens to be facing. Nonetheless, many people have social preferences that support conditionally cooperative behavior (Fehr & Fischbacher, 2004; Fischbacher, Gaechter, & Fehr, 2001). These people are willing to cooperate conditional on a sufficiently strong belief that others will also cooperate. For these people, social preferences transform a game that is nominally a social dilemma into a coordination game, and so beliefs about the propensity of others to cooperate can be decisive (Bowles, 2004, Camerer & Fehr, 2006).

In sum, beliefs about others play a fundamental role in diverse social settings. They can determine, among countless other phenomena, if an employee works hard on a team project (Bowles, 2004), when a Bolivian driver switches from one side of the road to the other (Camerer, 2003), which Ethiopian pastoralists conserve their natural resources (Rustagi, Engel, & Kosfeld, 2010), if Sudanese families circumcise their daughters (Mackie, 1996), and whether a customer asks for lagniappe at the local fruit stand. In many cases, beliefs are based on some kind of mutual history together because people have interacted repeatedly or they know they share some relevant cultural background. What, however, does someone do given little or no appropriate experience? What does someone think, and by extension

* Corresponding authors. Sonja Vogt, Charles Efferson, and Ernst Fehr are to be contacted at Department of Economics, University of Zurich, Blümlisalpstrasse 10, 8006 Zurich, Switzerland.
E-mail addresses: sonja.vogt@econ.uzh.ch (S. Vogt), charles.efferson@econ.uzh.ch (C. Efferson), ernst.fehr@econ.uzh.ch (E. Fehr).

1090-5138/$ – see front matter
http://dx.doi.org/10.1016/j.evolhumbehav.2013.03.003

how does someone behave, when interacting with a stranger or recent acquaintance? This paper focuses on the accuracy of beliefs about others under this kind of limited information. Specifically, we present results from an experiment that varied both the game people played and the amount of information available about these players to determine if and when, figuratively speaking, we can see inside.

Given limited information about someone, an especially simple approach would be to rely on beliefs about the distribution of choices made by randomly selected unfamiliar partners. Consider someone playing the stag hunt game in Table 1. The stag hunt game is a coordination game with two pure-strategy equilibria, namely both play stag or both play hare. If both play stag, both players receive a large payoff (60 in Table 1). Playing stag, however, involves a certain danger because, if one's partner plays hare, playing stag yields a low payoff (6 in Table 1). In contrast, if both players play hare, both receive an intermediate payoff (40 in Table 1), and playing hare involves little or no danger. Even if one's partner plays stag, one still receives an intermediate payoff by playing hare (40 in Table 1). Altogether, playing stag brings a large payoff if one coordinates and a small payoff if one miscoordinates. Playing hare brings an intermediate payoff if one coordinates and an intermediate payoff, possibly the same, if one miscoordinates. The key question about a stag hunt player is whether she will tolerate the danger of playing stag in order to support the potential for the large payoffs that come from coordinating on stag. Empirically, people vary in their tendencies to play stag because they vary in their willingness to tolerate this danger and in their beliefs about others (Camerer, 2003). Assume that, for whatever reason, a focal player believes a randomly selected partner will play stag with some probability. Further assume that, when paired with a stranger to play the game, our focal player simply asks herself if this probability is sufficiently large to play stag herself. The decision-making procedure in this case does not involve any information about the specific stranger at hand. It involves only an indiscriminate, unconditional belief about randomly selected partners.

For conditionally cooperative individuals, the same procedure could apply when playing the prisoner's dilemma game in Table 2. In this case, the relevant question about a prisoner's dilemma player is whether she will provide benefits for another even though this choice is always costly in material terms. As in the stag hunt game discussed above, assume a focal player believes a randomly selected individual will cooperate with some probability. When paired with a randomly selected stranger, the focal player asks herself if this probability is sufficiently large to cooperate herself. As before, this approach uses no information about the specific, unfamiliar partner who happens to be present. It depends only on disembodied beliefs about the population of potential partners.

We know, however, that people do not often rely on disembodied beliefs. Instead, they make snap judgments about others based on cursory contact and limited information. People discriminate based on ethnicity, gender, language, clothing, appearance, mannerisms, and many other traits that are readily observable (Carré, McCormick, & Mondloch, 2009; Dovidio, Glick, & Rudman, 2005; Fetchenhauer, Groothuis, & Pradel, 2010; Stirrat & Perrett, 2010; Willis & Todorov, 2006). Unlike the indiscriminate beliefs described above, beliefs are conditional on a partner's observable characteristics. Instead of ignoring the specific partner at hand, a decision maker somehow observes the person in front of her and makes a rapid assessment. One might try to assess, for example, whether another person can tolerate the danger of playing stag or if another person has the social preferences necessary to support conditional cooperation.

Table 1
The stag hunt game.

	Stag (#)	Hare (@)
Stag (#)	60, 60	6, 40
Hare (@)	40, 6	40, 40

For each outcome, payoffs are shown for the row player first and then the column player. For participants who played the game, three points were equivalent to one Euro. Although "Stag" and "Hare" are included here for clarity, players and raters only saw the arbitrary labels # and @.

Table 2
The prisoner's dilemma.

	Cooperate (#)	Defect (@)
Cooperate (#)	60, 60	20, 70
Defect (@)	70, 20	40, 40

For each outcome, payoffs are shown for the row player first and then the column player. For participants who played the game, three points were equivalent to one Euro. Although "Cooperate" and "Defect" are included here for clarity, players and raters only saw the arbitrary labels # and @.

Interestingly, several recent studies suggest that conditional beliefs based on observable traits could be accurate. As one important example, men with wide faces tend to be aggressive and untrustworthy, while independent raters tend to believe that men with wide faces are aggressive and untrustworthy (Carré & McCormick, 2008; Carré et al., 2009; Haselhuhn & Wong, 2012; Stirrat & Perrett, 2010). These studies do not show that beliefs about a specific man's behavior, conditional on observing a specific man's face, are accurate. Nonetheless, perceived behavioral tendencies and actual behavioral tendencies are statistically associated with facial width in the same way, which plainly suggests that conditional beliefs about individuals *could* be accurate. If conditional beliefs are accurate, they could dramatically improve the ability of decision makers to interact with others effectively. A decision maker, for example, could interact with another person but condition her choices on the person's type, where type is represented by observable characteristics. Alternatively, a decision maker could choose between interacting with the person or foregoing the exchange altogether to pursue some more promising use of her time. In either case, conditional beliefs and by extension conditional behavior could improve the expected outcome for the decision maker precisely because of an ability to rapidly draw accurate inferences about other people.

Evolutionary theory makes clear predictions about when inferences of this sort should be accurate. With social dilemmas, inferences should typically not be accurate, and the logic is compelling. A conditionally cooperative individual needs to identify those who will cooperate and those who will not in order to reduce the risk of exploitation (Henrich, 2004). If, however, a cooperative person has only limited information about a partner, how can she infer what kind of person this partner is? She can only make an accurate inference if the unobservable tendency to cooperate is reliably associated with an observable marker of some kind. If this is the case, cooperative individuals can condition their beliefs and their choices on the presence of the marker. This kind of system, however, will not be evolutionarily stable for arbitrary markers that have costs unrelated to behavior. Once we allow a mutation that produces the marker without the tendency to cooperate, the mutation in question will invade the population. Because we generally have no reason to preclude such a mutation (Henrich, 2004), we expect that readily observable traits will usually not be associated with an unobservable tendency to cooperate in social dilemmas (Dawkins, 1976; Efferson & Vogt, 2013). Consequently, accurate inferences under limited information will not be possible. Intuitively, individuals do not have a shared interest in accurate information. If cooperative individuals use observable traits as a basis for cooperating conditionally, material incentives strongly favor defectors who mimic cooperators and trick them into cooperating. Once we allow such a masquerade, it flourishes immediately, reduces the accuracy of conditional beliefs, and eliminates the advantages of marker-based conditional cooperation.

Coordination games are very different because players have a shared interest in accurate information and coordinated choices. If a population includes individuals who tend to play different behaviors in a coordination game, deceiving others brings little or no advantage. Although in some coordination games everyone may not agree about where to coordinate, everyone does have a shared interest in coordinating. To continue with our stag hunt example, some players may expect or prefer to coordinate on stag. Others, in contrast, may be unwilling to tolerate the potential for the big material loss (e.g. an unsuccessful hunt) that can occur when playing stag. These players may expect to coordinate on hare. Because players can vary in terms of their expectations or their preferences over material outcomes, they can vary in terms of whether they play stag or hare (Camerer, 2003). All players, however, prefer coordinating to miscoordinating. Consequently, the incentives to misrepresent one's likely behavior in the near future are much less than in social dilemmas. This means that arbitrary observable markers can be dynamically stable indicators of behavioral tendencies in coordination games. Even more strongly, markers that are initially meaningless can acquire meaning endogenously because they help people draw accurate inferences about each other. This kind of evolutionary process works precisely because people have a shared interest in coordination and the accurate information it requires. Ex post, those who play one equilibrium strategy can separate themselves from those who play another equilibrium strategy (Efferson, Lalive, & Fehr, 2008; McElreath, Boyd, & Richerson, 2003).

The upshot is the following. In a social dilemma, readily available information about how a person will behave should often be suspect, and inferences about others based on limited information should only produce accuracy rates at chance. In a coordination game, in contrast, our inclinations should often be written all over our faces. To test these predictions, we conducted an experimental study that directly addresses inferential accuracy about others under limited information. Specifically, one group of subjects played one of two strategic games, either the coordination game in Table 1 or the social dilemma in Table 2. We call these subjects "players." Subsequently, a second group of subjects watched short videos of these players and guessed their choices in the games. We call these subjects "raters." We predicted that raters would not be able to accurately guess the choices of social dilemma players. This follows from the logic, outlined above, that conspicuous markers of underlying behavioral tendencies should typically not be stable in a social dilemma. Thus, the raters in our experiment, who had only brief exposure to the social dilemma players via the videos we showed them, should not have been able to accurately guess player choices. In contrast, we predicted that raters would be able to accurately guess the choices of coordination game players. This prediction arises from the fact that arbitrary observable traits readily evolve to serve as stable markers of behavioral tendencies in coordination games. Observable traits, whether they evolve genetically or culturally, can acquire and retain meaning because everyone has some shared interest in accurate information about others. This shared interest can be especially critical if people vary in terms of their unobservable preferences over material outcomes or if they come from historically separated sub-populations (Efferson et al., 2008; McElreath et al., 2003). If an evolutionary process occurs under circumstances of this sort, it implies scope for inferential accuracy regardless of what the markers actually are in practice and regardless of whether people are fully aware of how they use them. With one important caveat, our data support none of the above predictions.

2. Experimental methods

Our experiment consisted of two parts (electronic supplementary material). For the first part in Konstanz, Germany, we video recorded subjects individually for 60 s and then had them play one of two games. Videos of this sort are called "thin slices" because they provide brief and relatively controlled access to the personality and characteristics of the person in the video (Ambady & Rosenthal, 1992). In the vast majority of the thin slices we recorded, subjects discussed their families, work, their studies at the university, and what they like to do in their free time. A handful of subjects described what they had done earlier in the day. One woman enthusiastically summarized her recent trip to India, and one man counted the chairs in the room and commented on the impassive experimenter (S.V.) behind the camera. After recording thin slices for all participants in an experimental session, participants played either the stag hunt game in Table 1 or the prisoner's dilemma game in Table 2. For the second part of the experiment, another group of participants in Munich, Germany, served as raters. These raters viewed thin slices of either stag hunt players or prisoner's dilemma players and then guessed the choices of these players in the relevant game. In addition, raters viewed the videos either with the sound on or with the sound off. As a result, the information available to raters varied because they either could or could not hear what the players in the thin slices were saying. This allowed us to see if an increase in the information available would lead to an improvement in rater accuracy. Altogether, our experiment implemented a 2×2, between-subjects design in which we varied both the game played by players and the amount of information available to raters.

In addition, we also ran separate sessions to measure the mean attractiveness of each player averaged over several independent participants whose only task was to evaluate player attractiveness (electronic supplementary material). Because the timbre of one's voice might affect perceived attractiveness, we ran a session with the sound of the thin slices off and a session with the sound on. This resulted in two mean attractiveness ratings per player, and these variables appear below as important controls in several analyses.

Finally, as detailed in the electronic supplementary material, we made a number of design choices to isolate and compare accuracy rates stemming from the thin slices themselves and the two games players played. First, for both games we used the same labeling system for the possible choices, and the labels used have no particular meaning or natural ordering (Tables 1 and 2). Second, we independently randomized the spatial location of inputs on the computer screen for each player and each rater. Together these two design choices meant that raters could not have an artificially inflated accuracy rate because both players and raters shared the same psychological focus on a specific label or a specific location on the input screens. Third, for each of the two games, we randomly sampled 30 players to show to raters subject to the constraint that the distribution of choices among these players would be uniform. As a result, raters viewed 30 thin slices of stag hunt players, 15 of whom chose stag and 15 hare. Similarly, raters viewed 30 thin slices of prisoner's dilemma players, 15 of whom chose cooperate and 15 defect. Raters knew they would be presented with a uniform distribution of choices, but they did not know how many thin slices they would view in total. This is how we controlled rater beliefs about player behavior prior to viewing a thin slice, and this is how we held these prior beliefs constant regardless of whether the rater viewed stag hunt players or prisoner's dilemma players. Controlling prior beliefs in this way is essential when comparing accuracy rates across the two games, and overall it was a key part of our strategy for isolating any effects associated with information in the thin slices. More generally, our design choices eliminated the possibility that observed accuracy rates might reflect unwanted experimental artifacts (electronic supplementary material). Altogether, 36 raters viewed stag hunt players with the sound off (1080 observations), 36 raters viewed stag hunt players with the sound on (1080 observations), 35 raters viewed prisoner's dilemma players with the sound off (1050 observations), and 36 raters viewed prisoner's dilemma players with the sound on (1080 observations). When modeling rater guesses, we control for multiple observations per rater by clustering

on rater (electronic supplementary material, available on the journal's website at www.ehbonline.org).

3. The use of thin slices

Before turning to the results, we would like to address a crucial methodological issue. Namely, when information about a person can take so many different forms, and when communication can occur in so many different ways, why should a researcher use thin slices? We see at least three compelling reasons. First, thin slices carry an extensive empirical precedent. Past research has shown that people can use thin slices to draw accurate inferences about others in a wide variety of domains, including marital happiness, sexual orientation, intelligence, socioeconomic status, and altruism (Ambady & Rosenthal, 1992; Ambady, Hallahan, & Conner, 1999; Borkenau, Mauer, Riemann, Spinath, & Angleitner, 2004; Fetchenhauer et al., 2010; Kraus & Keltner, 2009). As a result, previous research suggests that, for those examining how people draw inferences about others given limited information, thin slices offer an excellent place to start.

Second, thin slices represent a useful balance between experimental control and external validity. On the one hand, we can imagine a procedure in which the experimenter places a participant in a situation with precisely two possible behaviors, and the experimenter further requires the participant to choose one of two predetermined messages communicating the participant's intended behavior to some unknown person. This method offers complete control for the experimenter, but its similarity to social interactions outside the lab is arguably limited. On the other hand, we can imagine an alternative procedure in which the experimenter tells two participants, say a player and a rater, to go off and get to know each other for as long as they desire. When they are ready to continue, they can call the experimenter's mobile phone, then everyone will rendezvous in the lab and proceed with the study. Communication between the two participants in this case is extremely similar to communication outside the lab, but the scientist has no control of any kind over what happens.

Thin slices stand between these two extremes. They offer perfect control over the amount of time available for communicating. In addition, by decomposing a thin slice into an audio recording and a video recording, thin slices offer considerable control over the extent to which communication is verbal versus visual. Given that verbal language plausibly evolved from a human social psychology rooted in non-verbal communication (Tomasello, 2008), we can expect both types of communication to be important. All in all, thin slices provide an extremely useful method for admitting the subtleties of natural communication without ceding control as a researcher.

Finally, when thin slices are recorded, as ours were, before participants have a detailed knowledge of the upcoming social interaction, they capture the non-obvious nature of much communication. Specifically, games like the prisoner's dilemma and stag hunt game are abstract representations of broad classes of social interaction in which individuals make choices that affect others. Because of these external effects, social norms often play a strong role in governing behavior (Bowles, 2004; Bowles & Gintis, 2011). Directly and efficiently communicating a relevant social norm, however, is often not realistic. Social interactions do not always come with handy labels like prisoner's dilemma and stag hunt that make the relevant norm immediately obvious. In addition, groups may differ in terms of social norms but not fully realize, because of limited historical contact, exactly how they differ (McElreath et al., 2003). To make matters even more complicated, individuals can have different identities and roles in society that require different norms based on which identity is most active at a given point in time (Akerlof & Kranton, 2010; Benjamin, Choi, & Strickland, 2010). Put all these complexities together, and what a person must communicate may not always be obvious. In these cases, selection should create pressure for people who are different but have a shared interest in accurate information to somehow mark and essentialize group identity (Efferson et al., 2008; Gil-White, 2001; McElreath et al., 2003). This would allow people to efficiently draw statistically reliable inferences about others without having to rely exclusively on verbal communication to identify the relevant normative domain and negotiate any differences among the actors.

4. Results

We first present results for raters who viewed stag hunt players. A total of 45 players played the stag hunt game. Of these, 21 played # (stag), the choice associated with the payoff-dominant equilibrium. Our random sample of players, subject to a uniform distribution of choices, resulted in a sample of 10 men, six of whom played stag, and 20 women, nine of whom played stag. When the sound was off for the thin slices of these 30 players, raters did not guess player choices above chance. Specifically, over all guesses the proportion correct was 0.497, and the 95% robust confidence interval clustered on rater is [0.472, 0.522]. When the sound was on, raters were also not above chance, with an overall accuracy rate of 0.514 and a 95% robust confidence interval clustered on rater of [0.482, 0.545]. In addition, a probit regression of accuracy as a function of the four treatments also shows that the increase in accuracy when the sound was on is not significant and that rater accuracy was not above chance in either of the stag hunt treatments (Table 3).

Although rater accuracy was not above chance when viewing thin slices of stag hunt players, rater guesses may still have varied systematically in some way. To see if this was so, and in particular to see if rater guesses varied according to some attribute of players or raters, we conducted a large model selection exercise (electronic supplementary material) using information theoretic criteria (Burnham & Anderson, 2002). This exercise produced the following robust result. When viewing thin slices of stag hunt players, with the sound either off or on, the attractiveness of the player was the key variable associated with rater guesses (electronic supplementary material, Tables S2 and S5, available on the journal's website at www.ehbonline). In particular, raters guessed # (stag) with a higher probability for more attractive players. Although the exact size of the effect varied some according to model specification and whether the sound was on or off, altogether it was robust, positive, and highly significant (probit regressions with robust standard errors clustered on rater, $p \leq 0.002$, electronic supplementary material, Tables S2 – S7, available on the journal's website at www.ehbonline). Nonetheless, in spite of the fact that raters guessed # more often for attractive players, raters did not use and indeed could not have used this information to

Table 3
Accuracy as a function of treatment.

Parameter	Estimate	Robust Std. Error	p-value
Intercept	−0.007	0.031	0.820
Sound	0.042	0.049	0.391
PD	0.100	0.044	0.024
Sound × PD	0.031	0.072	0.670

The accuracy of 4290 guesses is modeled using a probit regression with robust standard errors clustered on 143 raters. Raters guessed the choices of players who played either the stag hunt game or the prisoner's dilemma (PD), and raters viewed thin slices with the sound either off or on (Sound). Because the intercept is not significant, accuracy was at chance in the stag hunt treatment with the sound off. Accuracy was also at chance with the sound on ($\hat{\beta}_{\text{Int}} + \hat{\beta}_{\text{Sound}} = 0.035, p = 0.358$). Finally, because the estimate for sound is not significant, turning the sound on did not produce a significant increase in accuracy. For the prisoner's dilemma, accuracy was significantly above chance with the sound off ($\hat{\beta}_{\text{Int}} + \hat{\beta}_{\text{PD}} = 0.093$, p = 0.004) and with the sound on ($\hat{\beta}_{\text{Int}} + \hat{\beta}_{\text{Sound}} + \hat{\beta}_{\text{PD}} + \hat{\beta}_{\text{Sound} \times \text{PD}} = 0.166, p < 0.001$). As the interaction term shows, however, turning the sound on did not produce a significant increase in accuracy. Overall, χ^2 (3) = 12.80, $p = 0.005$, and the pseudo-$R^2 = 0.002$. The limited explanatory power is due to the fact that accuracy, even when significant, never varied far from chance for any of the treatments.

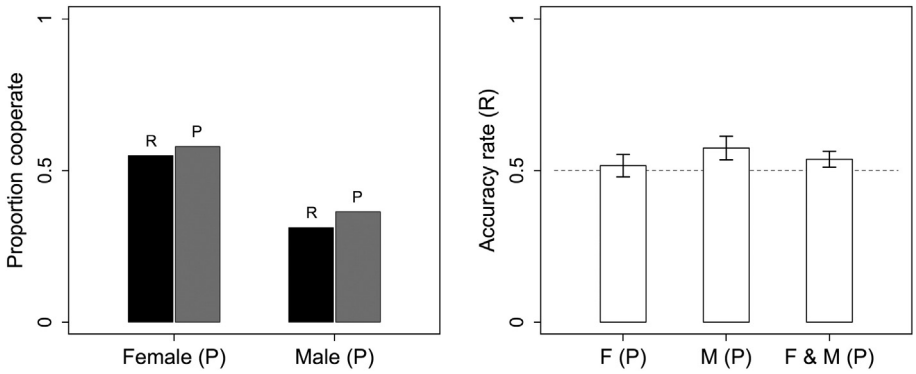

Fig. 1. Player (P) behavior and rater (R) guesses separated by the sex of the player when the sound of the thin slices was off. The left panel shows the proportion of times that raters (black bars) guessed cooperate and players actually cooperated (grey bars) for female players (P) and male players (P). When comparing female players to male players, we see similar increases in the rate at which players actually chose to cooperate and the rate at which raters guessed cooperation. This increase is highly significant for the 1050 rater guesses shown (probit regression with robust standard errors clustered on rater, coefficient on female is 0.614, $p < 0.001$) but not for the 30 player choices (coefficient on female is 0.548, $p = 0.256$). The right panel shows the accuracy of raters for female players (F(P)), male players (M(P)), and all players (F & M (P)) with 95% confidence intervals clustered on rater. For female players, the confidence interval is [0.479, 0.553]. For male players, it is [0.535, 0.613], and over all players it is [0.511, 0.563].

improve the accuracy of their guesses. Given attractiveness levels based on thin slices with the sound off (see Fig. S2, available on the journal's website at www.ehbonline.org and associated probit regressions in the electronic supplementary material), player choices were not related to player attractiveness (probit regression, $p = 0.467$), rater accuracy was not significantly related to player attractiveness (probit regression, $p = 0.496$), and estimated rater accuracy was not different from chance for either extreme levels of unattractiveness (probit regression, $p = 0.559$) or attractiveness (probit regression, $p = 0.462$) outside our sample of players. Similarly, using attractiveness levels from thin slices with the sound on (see Fig. S3, available on the journal's website at www.ehbonline and associated probit regressions in the electronic supplementary material), player behavior was not significantly related to player attractiveness (probit regression, $p = 0.179$), rater accuracy was not significantly related to player attractiveness (probit regression, $p = 0.568$), and estimated rater accuracy was not significantly different from chance for the two most extreme levels of attractiveness outside our sample (probit regression, $p = 0.737$ and $p = 0.422$).

A total of 52 players played the prisoner's dilemma, and 15 of them chose # (cooperate). Our random sample of players, subject to a uniform distribution of choices, consisted of 11 males, four of whom cooperated, and 19 females, 11 of whom cooperated. When guessing the behavior of these 30 players, raters were significantly above chance when the sound of the thin slices was both off and on. With the sound off, the proportion of accurate guesses was 0.537 with a 95% robust confidence interval clustered on rater of [0.511, 0.563]. With the sound on, raters guessed correctly at a rate of 0.566, and the 95% robust confidence interval clustered on rater is [0.532, 0.599]. A probit regression of accuracy as a function of treatment also indicates that accuracy was above chance in both of the prisoner's dilemma treatments (Table 3). The increase in accuracy that followed from turning the sound on, however, is not significant (Table 3).

To see if rater guesses varied systematically in the prisoner's dilemma treatments, we conducted another large model selection exercise (electronic supplementary material). This produced clear and robust results. Namely, regardless of whether the sound of the thin slice was on or off, the sex of the player in the thin slice was a critical variable related to rater guesses (electronic supplementary material, Tables S8 and S11, available on the journal's website at www.ehbonline). In particular, raters guessed that females cooperated more than males, and across multiple regressions identified by the model selection criterion this effect is robustly significant (probit regressions with robust standard errors clustered on rater, $p \leq 0.01$, electronic supplementary material, Tables S8–S13, available on the journal's website at www.ehbonline). As mentioned above, the females in our sample of players did cooperate at a higher rate than the males, but this difference is not significant (probit regression, $p = 0.256$, Figs. 1 and 2). Nonetheless, raters were able to use the thin slices to get above chance with their guesses. When

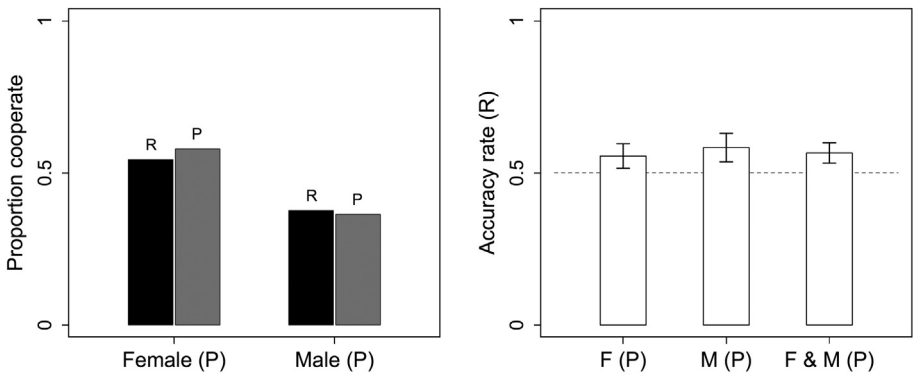

Fig. 2. Player (P) behavior and rater (R) guesses separated by the sex of the player when the sound of the thin slices was on. The left panel shows the proportion of times that raters (black bars) guessed cooperate and players actually cooperated (grey bars) for female players (P) and male players (P). When comparing female players to male players, we see similar increases in the rate at which players actually chose to cooperate and the rate at which raters guessed cooperation. This increase is highly significant for the 1080 rater guesses shown (probit regression with robust standard errors clustered on rater, coefficient on female is 0.425, $p = 0.001$) but not for the 30 player choices (coefficient on female is 0.548, $p = 0.256$). The right panel shows the accuracy of raters for female players (F(P)), male players (M(P)), and all players (F & M (P)) with 95% confidence intervals clustered on rater. For female players, the confidence interval is [0.515, 0.596]. For male players, it is [0.537, 0.630], and over all players it is [0.532, 0.599].

the sound was off, both the accuracy rate for male players and the overall accuracy rate are significant in the sense that the 95% robust confidence intervals clustered on rater did not span 0.5 (Fig. 1). When the sound was on, the accuracy rates for both male and female players are significant in this sense, and the overall accuracy rate is thus also necessarily significant (Fig. 2).

Regression analyses (electronic supplementary material) also identified inferential accuracy in one additional way. When restricting attention to raters who viewed thin slices of prisoner's dilemma players with the sound on, rater guesses and player choices are significantly and positively related (probit regressions with standard errors clustered on rater, $p = 0.001$, electronic supplementary material, Tables S11–S13, available on the journal's website at www.ehbonline). As discussed above, turning the sound on did not produce a significant *increase* in the accuracy of raters viewing prisoner's dilemma players. Nonetheless, rating prisoner's dilemma players with the sound on yielded the highest accuracy rate over all treatments, and this fact is captured by a significant relationship between rater guesses and player choices in this treatment.

Finally, over all four treatments we have little or no evidence that observed accuracy rates reflect a heterogeneous mix of raters with some raters guessing accurately and others guessing inaccurately (electronic supplementary material). Instead, raters seem to have been fairly homogeneous. To show this, for each treatment, $\forall\, k \in \{0, 1, \ldots, 30\}$, we calculated the expected number of raters with k correct guesses under the assumption that all raters are identical. We then compared these theoretical distributions to the observed distributions for each of the four treatments. Goodness-of-fit tests like chi-squared or the G test are not valid here because the expected numbers of raters for many outcomes are extremely small (electronic supplementary material). Nonetheless, visually inspecting the graphs (electronic supplementary material, Fig. S1, available on the journal's website at www.ehbonline) clearly shows that rater heterogeneity, if it exists at all, can have at most a minor role in our data. Moreover, we also used a probit regression to analyze rater accuracy as a function of treatment and individual-level variables that control for the gender, age, empathic concern, and perspective-taking ability of each rater (electronic supplementary material, Table S1, available on the journal's website at www.ehbonline). As in Table 3, the effect for the prisoner's dilemma dummy is significant, but none of the individual-level controls are significant. This finding also indicates that heterogeneity among raters in terms of accuracy plays little or no role in our data.

5. Discussion and conclusion

In contrast to the prediction that inferential accuracy should be higher for coordination games than for social dilemmas, we found that raters guessed the choices of prisoner's dilemma players more accurately than the choices of stag hunt players. Moreover, in contrast to the prediction that rater accuracy should not be above chance when guessing behavior in social dilemmas, we found accuracy rates significantly above chance when raters viewed thin slices of prisoner's dilemma players. This was true for thin slices with the sound off and for thin slices with the sound on. When the sound was off, rater accuracy was driven primarily by guessing the choices of male players. When the sound was on, raters were above chance for both male and female players.

Although statistically significant, how meaningful are the accuracy rates we observed with respect to the evolution of cooperation? To get some grip on this question, we developed two different models, one based on conditional behavior and the other based on conditional group formation (electronic supplementary material). Both of these models include predictive accuracy, which we call q, as a key parameter. Importantly, because we treat q as a parameter, we do not address the evolutionary dynamics of inferential accuracy. We simply posit an accuracy rate of q and follow the consequences. In particular, we take our observed accuracy rate from the prisoner's dilemma treatment with the sound on as a benchmark value (i.e. $q = 0.566$). This is a best-case scenario for accurate inferences to support the evolution of cooperation because it is the highest overall accuracy rate we observed. Given this accuracy rate, we identify the properties a prisoner's dilemma must have for cooperation to evolve when people can, metaphorically, see inside at a rate of $q = 0.566$ and thus reduce the risk of exploitation.

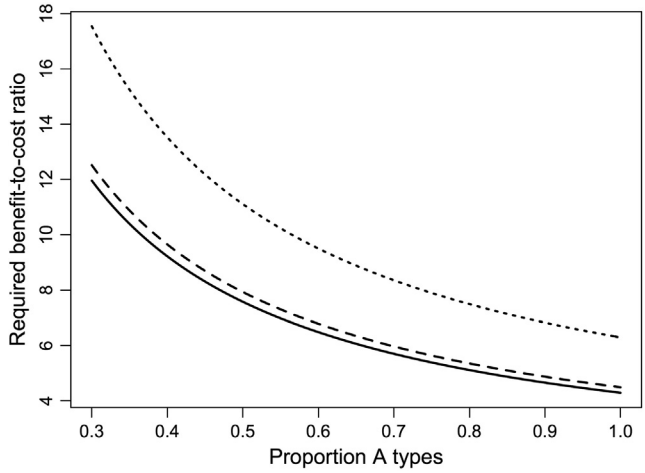

Fig. 3. The minimum benefit-to-cost ratio required for cooperation to evolve given the inferential accuracy we observed in the prisoner's dilemma treatment with the sound on ($q = 0.566$). The solid line shows the minimum required benefit-to-cost ratio under the conditional cooperation model as a function of the proportion of A types in the population. The dashed line shows the minimum required benefit-to-cost ratio under the conditional group formation model when the outside option involves a relatively low payoff, and the dotted line shows the same when the outside option involves a relatively high payoff. See the electronic supplementary material for additional details.

For both models, the population consists of two types of individual, A and N. In the conditional behavior case, pairs are formed randomly to play a simultaneous prisoner's dilemma. Each A type guesses the type of her partner, and she cooperates if she thinks her partner is also an A. Otherwise she defects. A guess is accurate with probability q. N types defect unconditionally. In the conditional group formation case, pairs are formed randomly. Each individual guesses the type of her partner, and these guesses are accurate with probability q. If an A type plays, she cooperates. If an N type plays, she defects. A pair plays only if both individuals agree to play. Individuals of both types only agree to play if they think they are paired with an A type. If one or both players refuse to play, each player gets some benefit associated with their best outside option.

The models show that the inferential accuracy we observed, though statistically significant, is unlikely to be evolutionarily meaningful. Under the conditional behavior model, the benefit-to-cost ratio of cooperation must exceed approximately 4.29 just to render A resistant to invasion by N (Fig. 3). If half of the population consists of A individuals, the minimum ratio for A to evolve is 7.58, and minimum ratios increase rapidly from there for populations with a majority of N types (Fig. 3). If individuals form groups conditionally, the situation is even worse for cooperation. In particular, conditional group formation is equivalent to conditional behavior when the outside option brings no benefits (electronic supplementary material). As the outside option improves, conditions for the evolution of cooperation deteriorate in the sense that the required benefit-to-cost ratio increases (electronic supplementary material). When outside options are good, in particular, for virtually any distribution of types the required benefit-to-cost ratio is unreasonably high for A to evolve under conditional group formation and the inferential accuracy we observed in our experiment (Fig. 3).

Earlier, we argued that inferential accuracy should be at chance for social dilemmas and significantly above chance for coordination games. We also prefigured that, with one caveat, our data support neither of these predictions. Specifically, in contrast to the predictions, inferential accuracy in our experiment was at chance for the stag hunt and significantly above chance for the prisoner's dilemma. The caveat is the following. Although accuracy was statistically above chance for the prisoner's dilemma, our calibration exercise shows that accuracy was probably not high enough to be evolutionarily meaningful. Accurately identifying cooperative tendencies with a probability of 0.566 can generate some assortment, but not very much. As a result, the limited information represented by the thin slices we recorded could only support the evolution of cooperation under especially large benefit-to-cost ratios. In this evolutionary sense, raters of prisoner's dilemma players, like raters of stag hunt players, were *effectively* at chance in our experiment.

In spite of the fact that accuracy was poor, rater guesses did vary systematically with the attractiveness and sex of players. This finding suggests that raters were trying to use the information in the thin slices to draw accurate inferences; they just could not do so. Presumably, with increasing amounts of information inferences would eventually be meaningfully accurate. Both the amount and the type of information about another person can vary. At one extreme, we have the anonymous interactions that typify economic experiments (Camerer, 2003). At the other extreme, we can imagine two people who have known each other for years and have a close personal relationship. At some point between these extremes, the amount and type of information available should allow one person to accurately predict the behavior of the other person in a particular type of social interaction. If either the amount of information is inadequate or the type of information is inappropriate, accuracy will not be above chance. In our study, for example, rater accuracy might have been at chance for stag hunt players because the thin slices we recorded did not capture the right kind of information. Precisely because any set of arbitrary markers can evolve to serve as coordination devices (Efferson et al., 2008; McElreath et al., 2003), the space of markers that can potentially serve this role is very large indeed. In effect, from symbols of group affiliation to non-verbal and verbal languages, many different kinds of language can be used to convey the information players need to coordinate. Moreover, raters may or may not use information effectively. In-group favoritism, parochialism, and associated prejudices may interact with information a person has about specific individuals to reduce or increase the probability of accurately assessing another's intentions. The larger scientific task is to delineate, for a given type of social interaction, how much information people require and what kind of information they require. We found that raters tended to guess that attractive people play stag and that women cooperate. These patterns were robust, but neither yielded accuracy rates high enough to be evolutionarily meaningful. Given 60-s thin slices, people try to see inside, but they cannot.

Supplementary Materials

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.evolhumbehav.2013.03.003.

Acknowledgments

We would like to acknowledge the generous support of the Swiss National Science Foundation (Grant No. 100014-130127/1, "The Social Dynamics of Normative Behavior") and the University of Zurich Research Priority Program, "Foundations of Human Social Behavior — Altruism versus Egoism." We also thank Paul Seabright for helpful comments on an earlier version of this paper. Finally, we extend a special thank you to our colleagues at MELESSA in Munich and the LakeLab in Konstanz for all the help associated with using their facilities.

References

Akerlof, G. A., & Kranton, R. E. (2010). Identity economics: How our identities shape our work, wages, and well-being. Princeton: Princeton University Press.
Ambady, N., Hallahan, M., & Conner, B. (1999). Accuracy of judgments of sexual orientation from thin slices of behavior. *Journal of Personality and Social Psychology*, *77*(3), 538–547.
Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*(2), 256–274.
Benjamin, D. J., Choi, J. J., & Strickland, A. J. (2010). Social identity and preferences. *American Economic Review*, *100*(4), 1913–1928.
Borkenau, P., Mauer, N., Riemann, R., Spinath, F., & Angleitner, A. (2004). Thin slices of behavior as cues of personality and intelligence. *Journal of Personality and Social Psychology*, *86*(4), 599–614.
Bowles, S. (2004). Microeconomics: Behavior, institutions, and evolution. New York: Russell Sage.
Bowles, S., & Gintis, H. (2011). A cooperative species: Human reciprocity and its evolution. Princeton: Princeton University Press.
Burnham, K. P., & Anderson, D. R. (2002). Model selection and multimodel inference: A practical information-theoretic approach (2nd ed.). New York: Springer-Verlag.
Camerer, C. F. (2003). Behavioral game theory: Experiments in strategic interaction. Princeton: Princeton University Press.
Camerer, C. F., & Fehr, E. (2006). When does "economic man" dominate social behavior? *Science*, *311*, 47–52.
Carré, J. M., & McCormick, C. M. (2008). In your face: Facial metrics predict aggressive behaviour in the laboratory and in varsity and professional hockey players. *Proceedings of the Royal Society B*, *275*(1651), 2651–2656.
Carré, J. M., McCormick, C. M., & Mondloch, C. J. (2009). Facial structure is a reliable cue of aggressive behavior. *Psychological Science*, *20*(10), 1194–1198.
Dawkins, R. (1976). The selfish gene. Oxford: Oxford University Press.
Dovidio, J. F., Glick, P., & Rudman, L. A. (Eds.). (2005). *On the nature of prejudice: Fifty years after Allport*. Oxford: Blackwell Publishing.
Efferson, C., Lalive, R., & Fehr, E. (2008). The coevolution of cultural groups and ingroup favoritism. *Science*, *321*(5897), 1844–1849.
Efferson, C., & Vogt, S. (2013). Viewing men's faces does not lead to accurate predictions of trustworthiness. *Scientific Reports*, *3*(1047), 1–7.
Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*(4), 185–190.
Fetchenhauer, D., Groothuis, T., & Pradel, J. (2010). Not only states but traits — humans can identify permanent altruistic dispositions in 20 s. *Evolution and Human Behavior*, *31*(2), 80–86.
Fischbacher, U., Gaechter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economic Letters*, *71*(3), 397–404.
Gil-White, F. J. (2001). Are ethnic groups biological 'species' to the human brain? Essentialism in our cognition of some social categories. *Current Anthropology*, *42*(4), 515–554.
Haselhuhn, M. P., & Wong, E. M. (2012). Bad to the bone: Facial structure predicts unethical behavior. *Proceedings of the Royal Society B*, *279*(1728), 571–576.
Henrich, J. (2004). Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior and Organization*, *53*(1), 3–35.
Kraus, M. W., & Keltner, D. (2009). Signs of socioeconomic status: A thin-slicing approach. *Psychological Science*, *20*(1), 99–106.
Mackie, G. (1996). Ending footbinding and infibulation: A convention account. *American Sociological Review*, *61*, 999–1017.
McElreath, R., Boyd, R., & Richerson, P. J. (2003). Shared norms and the evolution of ethnic markers. *Current Anthropology*, *44*(1), 122–129.
Rustagi, D., Engel, S., & Kosfeld, M. (2010). Conditional cooperation and costly monitoring explain success in forest commons management. *Science*, *330*, 961–965.
Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust. *Psychological Science*, *21*(3), 349–354.
Tomasello, M. (2008). Origins of human communication. Cambridge: The MIT Press.
Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-Ms exposure to a face. *Psychological Science*, *17*(7), 592–598.
Young, H. P. (1996). The economics of convention. *The Journal of Economic Perspectives*, *10*(2), 105–122.